TAXI DEMAND PREDICTION USING ENSEMBLE MODEL AND TAXI GPS DATA ANALYSIS

BY

UKRISH VANICHRUJEE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGINEERING (INFORMATION AND COMMUNICATION TECHNOLOGY FOR EMBEDDED SYSTEMS)

SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2017

Ref. code: 25605922040588APL
TAXI DEMAND PREDICTION USING ENSEMBLE MODEL AND TAXI GPS DATA ANALYSIS

BY

UKRISH VANICH RUJEE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGINEERING (INFORMATION AND COMMUNICATION TECHNOLOGY FOR EMBEDDED SYSTEMS)

SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2017

Ref. code: 25605922040588APL
TAXI DEMAND PREDICTION USING ENSEMBLE MODEL AND TAXI GPS
DATA ANALYSIS

A Thesis Presented

By
UKRISH VANICHRUJEE

Submitted to
Sirindhorn International Institute of Technology
Thammasat University
In partial fulfillment of the requirements for the degree of
MASTER OF ENGINEERING (INFORMATION AND COMMUNICATION
TECHNOLOGY FOR EMBEDDED SYSTEMS)

Approved as to style and content by

Advisor

(Asst. Prof. Dr. Teerayut Horanont)

Committee Member and
Chairperson of Examination Committee

(Dr. Wasan Pattara-Atikom)

Committee Member

(Prof. Dr. Thanaruk Theeramunkong)

Committee Member

(Assoc. Prof. Dr. Takahiro Shinozaki)

MAY 2018
Acknowledgements

This research is financially supported by Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology, Sirindhorn International Institute of technology, Thammasat University under the TAIST Tokyo Tech Program. I would like to express my sincere gratitude to my advisor Asst. Prof. Dr. Teerayut Horanont and all committees for their guidance throughout this research. Moreover, I would like to thank all members in AI lab for their helps. Lastly, I would like to express my deepest gratitude to my family for supporting me throughout this study.
Abstract

TAXI DEMAND PREDICTION USING ENSEMBLE MODEL AND INTERACTIVE DATA VISUALIZATION

by

URKISH VANICHRUJEE

Bachelor of Engineering in Computer Engineering, Thammasat University, 2014
Master of Engineering (Information and Communication Technology for Embedded System), Sirindhorn International Institute of Technology, Thammasat University, 2018

Taxis play an important role in urban transportation. There are also many growing businesses related to taxi service such as Grab taxi, Uber etc. Obtaining the demand for transportation in the future can provide many business advantages. Taxi GPS data can be a good source for mining the demand for transportation. Understanding the taxi demand data in the future gives an opportunity to organize the taxi fleet better. Even, there are some works proposed to predict the demand of taxi but there are few studies that consider the function of areas such as hospital area, department store area, residential area, and tourist attraction. One predictive model may fit with all types of area. In this study, we use a point of interest (POI) to match taxi demand with a place to study the taxi demand in the area with a different function. First, we investigate the best predictive models that can forecast demand of taxi hourly with 7 types of area function which are airport, residential area, department store, education area, hospital, subway and tourist attraction. The models that we selected for the experiment are long short-term memory (LSTM), gated recurrent unit (GRU) and extreme gradient boosting (XGBOOST). Then, we developed an ensemble model that can forecast the demand of taxi well with all types of area function by combining those predictive models. We build the models based on a real-world dataset generated by over 5,000 taxis in Bangkok, Thailand for 4 months. The result shows that the proposed ensemble model can achieve
the smallest error rate with sMAPE of 23.72%. it can outperform the other standalone models.

**Keywords**: Taxi Demand Prediction, Time Series, Neuron Networks, Spatio-temporal Data
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Signature Page</td>
<td>i</td>
</tr>
<tr>
<td></td>
<td>Acknowledgements</td>
<td>ii</td>
</tr>
<tr>
<td></td>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td></td>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td></td>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td></td>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.1 Background</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.2 Data visualization</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1.3 Machine learning models</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1.3.1 Recurrent Neural Networks (RNNs)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1.3.1.1 Long Short-Term Memory (LSTM)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1.3.1.2 Gated Recurrent Unit (GRU)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1.3.2 Extreme Gradient Boosting (XGBOOST)</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>1.4 Statement of problem</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>1.5 Purpose of study</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>1.6 Significant of study</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1.7 Contributions</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>1.8 Outline</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>Literature Review</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>Bangkok taxi dataset and Data Preprocessing</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>3.1 Bangkok taxi dataset</td>
<td>14</td>
</tr>
</tbody>
</table>

v
3.2 Data Preprocessing

3.2.1 Extracting Origin-Destination
3.2.2 Filtering invalid trips
3.2.3 Matching the pick-up event to POI
3.2.4 Calculating the aggregated demand in every hour
   for each point of interest

3.3 Basic statistics

4 Methodology for taxi demand prediction

4.1 Types of area function
4.2 Train/Test split
4.3 Weather data
4.4 Features
4.5 Selected Predictive Models
4.6 Normalization
4.7 Evaluation Metric
4.8 Parameter tuning for RNNs
  4.8.1 Number of hidden neurons
  4.8.2 Optimizer
  4.8.3 Number of hidden layer
4.9 Parameter tuning for XGBOOST
4.10 Ensemble model

5 Interactive taxi visualization dashboard, interesting insights
   and taxi demand pattern of 7 types of area function

5.1 Interactive taxi visualization dashboard
  5.1.1 Data preparing for visualization
5.2 Interesting insights
  5.2.1 Comparison of taxi demand patterns between top
     and ordinary taxi drivers
5.2.2 Can number of taxi trips imply an anomaly event? 30
5.2.3 Top 10 taxi demand districts in Bangkok 31
5.3 Taxi demand pattern of 7 types of area function 32
5.3.1 Airport 32
      5.3.1.1 Airport outgoing trips 32
      5.3.1.2 Airport incoming trips 35

6 Experiment Result 37
6.1 Optimal network architecture for recurrent neural networks 37
   6.1.1 Number of hidden neurons 37
   6.1.2 Optimization algorithms 38
   6.1.3 Number of hidden layers 38
   6.1.4 Final architecture of LSTM and GRU 38
6.2 parameters for the ensemble model 39
   6.2.1 Coefficient and lookback window 39
   6.2.2 Final architecture of the ensemble model 40
6.3 Performance of the final models 41
   6.3.1 Performance of the final models in 7 types of area function 42
   6.3.2 Comparison of predicted demand and actual demand 43
6.4 The peak hour, off-peak of incoming trips and outgoing trips in 7 types of area function 45

5 Conclusion 46

References 47

Appendix 51

Appendix A 51
### List of Tables

<table>
<thead>
<tr>
<th>Tables</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Specification of taxi trace</td>
<td>14</td>
</tr>
<tr>
<td>6.1 sMAPE result of LSTM and GRU with different number of hidden neurons</td>
<td>37</td>
</tr>
<tr>
<td>6.2 sMAPE result of LSTM and GRU with different optimization algorithm</td>
<td>38</td>
</tr>
<tr>
<td>6.3 sMAPE result of LSTM and GRU with different number of hidden layers</td>
<td>38</td>
</tr>
<tr>
<td>6.4 sMAPE result of the ensemble model with coefficients (0.7, 0.2, 0.1)</td>
<td>39</td>
</tr>
<tr>
<td>6.5 sMAPE result of the ensemble model with coefficients (0.9, 0.1, 0)</td>
<td>39</td>
</tr>
<tr>
<td>6.6 Average sMAPE of all models in all places</td>
<td>41</td>
</tr>
<tr>
<td>6.7 Average sMAPE of all models in 7 types of area function</td>
<td>42</td>
</tr>
<tr>
<td>6.8 The peak, off-peak hour of outgoing and incoming trips in 7 types of area function</td>
<td>42</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figures</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Taxi demand patterns in different area functions</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Heatmap of taxi demand in Bangkok, Thailand</td>
<td>2</td>
</tr>
<tr>
<td>1.3 The architecture of a recurrent network</td>
<td>3</td>
</tr>
<tr>
<td>1.4 An unrolled recurrent network</td>
<td>3</td>
</tr>
<tr>
<td>1.5 The structure of the repeating module in the LSTM</td>
<td>5</td>
</tr>
<tr>
<td>1.6 The structure of the repeating module in the GRU</td>
<td>6</td>
</tr>
<tr>
<td>1.7 An example of the ensemble of Gradient Boosting Trees</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Heatmap of taxi pickup in Bangkok</td>
<td>10</td>
</tr>
<tr>
<td>3.1 The overview of data processing</td>
<td>15</td>
</tr>
<tr>
<td>3.2 Origin-Destination Extracting</td>
<td>16</td>
</tr>
<tr>
<td>3.3 Matching the pick-up event to POI</td>
<td>16</td>
</tr>
<tr>
<td>3.4 Example of aggregated demand in 24 time slots</td>
<td>17</td>
</tr>
<tr>
<td>3.5 Heatmap of taxi demand per hour in Bangkok city</td>
<td>17</td>
</tr>
<tr>
<td>3.6 Distance of Bangkok taxi trips</td>
<td>18</td>
</tr>
<tr>
<td>3.7 Duration of Bangkok taxi trips</td>
<td>18</td>
</tr>
<tr>
<td>3.8 Average speed of Bangkok taxi trips</td>
<td>19</td>
</tr>
<tr>
<td>3.9 Bangkok Taxi demand by hour</td>
<td>19</td>
</tr>
<tr>
<td>4.1 Dataset splitting</td>
<td>21</td>
</tr>
<tr>
<td>4.2 Example of weather data</td>
<td>21</td>
</tr>
<tr>
<td>4.3 Rolling window</td>
<td>22</td>
</tr>
<tr>
<td>4.4 Architecture of the ensemble model</td>
<td>25</td>
</tr>
<tr>
<td>5.1 Example of the interactive visualization dashboard</td>
<td>27</td>
</tr>
<tr>
<td>5.2 Example of a taxi data in JSON format</td>
<td>28</td>
</tr>
<tr>
<td>5.3 Comparison of taxi demand patterns between top and ordinary drivers</td>
<td>29</td>
</tr>
<tr>
<td>using the taxi visualization dashboard</td>
<td></td>
</tr>
<tr>
<td>5.4 Number of top and ordinary taxi driver trips by hour</td>
<td>30</td>
</tr>
<tr>
<td>5.5 Number of taxi trips to Muang Thong Thani on February, 2016</td>
<td>30</td>
</tr>
<tr>
<td>5.6 Top 10 taxi demand districts for different time periods in Bangkok</td>
<td>31</td>
</tr>
<tr>
<td>5.7 Heatmap of taxi demand at airport areas</td>
<td>32</td>
</tr>
</tbody>
</table>
5.8 Taxi demand at airport area on different day of week 33
5.9 Taxi demand pattern on weekday and weekend from airport areas by hour 33
5.10 Distance, duration and average speed of outgoing trips from airport areas 34
5.11 Heatmap of taxi demand to airport areas 35
5.12 Taxi demand to airport areas with different day of week 35
5.13 Duration of Bangkok taxi trips 36
5.14 Distance, duration and average speed of trips to airport areas 36
6.1 Final architecture of the ensemble model 40
6.2 One-week taxi demand prediction at Don Mueang Airport 43
6.3 One-week taxi demand prediction at Chatuchak Subway station 43
6.4 One-week taxi demand prediction at Asiatique 44
Chapter 1
Introduction

1.1 Background

Where to find a passenger is one of the most important questions for all taxi drivers. The more time taxi driver spends on cruising a new passenger, the more fuel consumption and the less number of passengers can be picked-up. For inexperienced taxi drivers, they usually don’t know where to pick-up a new passenger since they have no experience about the demand of taxi over time and space. The information about taxi demand in the future can be used to guide both inexperienced and experienced drivers to catch up with the taxi demand in the city faster. Knowledge about demand of taxi in a specific area can also be used to create new business opportunities or new services. For example, the taxi pick-up events that occurred around hotels can be used to discover the area that hotel clients usually visit in the morning. The hotel can provide a shuttle bus to the route that passes that specific place to gain more customer satisfaction.

With Rich information from taxi GPS sensors, there are many researchers study on how to use this digital footprint to discover some knowledge that can help us improving the transportation system. Taxi Demand prediction is one of those intelligence transportation systems. The demand patterns in area with different function are unique. Therefore, a single predictive model may not fit all areas.
Figure 1.1 shows examples of the taxi demand patterns in different functions.

Figure 1.1 Taxi demand patterns in different area functions

1.2 Data Visualization

Data visualization is the techniques used to communicate data or information clearly and understandable to users. Dealing with big data, data visualization is a necessary process to show how we can play with the datasets. Some hidden anomalies may be discovered if we should the right data visualization technique.

Figure 1.2 Heat map of taxi demand in Bangkok, Thailand
1.3 Machine Learning models

1.3.1 Recurrent Neural Networks (RNNs)

Recurrent neural networks (RNN) are designed specifically to operate over sequential data. It allows signal to travel both forward and backward. It has loops for internal connections among hidden neurons.

Figure 1.3: The architecture of a recurrent network

Figure 1.4: An unrolled recurrent network

Figure 1.3 shows the architecture of RNN. Given an input of time series $X = \{x_1, x_2, x_3, \ldots, x_t\}$, the RNN computes the hidden state sequence $H = \{h_1, h_2, h_3, \ldots, h_t\}$ and output sequence $Y = \{y_1, y_2, y_3, \ldots, y_t\}$ iteratively using the following equations.

\[
\begin{align*}
    h_t &= f(W_{hx}x_t + W_{hh}h_{t-1} + b_h) \\
    y_t &= g(W_{yh}h_t + b_y)
\end{align*}
\]
In (1) and (2), $W_{hx}$, $W_{hh}$ and $W_{yh}$ denote the input-to-hidden weight matrix, the hidden-to-hidden weight matrix and the hidden-to-output weight matrix respectively. The vector $b_h$ and $b_y$ are the bias of the hidden layer and the output layer respectively. $f(\cdot)$ and $g(\cdot)$ are the activation for the hidden layer and output layer respectively. The hidden state of time step $t$ will be passed to the hidden state of time step $t + 1$.

However, the conventional recurrent networks suffer from the gradient vanishing problem when it works with multi-step dependencies. Therefore, long short-term memory networks (LSTM) and Gated recurrent units (GRUs) were invented later. They are explicitly designed to avoid the gradient vanishing problem in the traditional recurrent neural networks.

1.3.1.1 Long Short-Term Memory (LSTM)

Long short-term memory [13] or LSTM is one of the most popular artificial neural networks currently because it can achieve high accuracy result in many kinds of research such as Natural language processing (NLP) and time series researches. The repeating module in LSTM has more complicated structure than the traditional RNNs. It contains a cell state, a forget gate, an input a gate and an output gate. Figure 1.5 shows the structure of the repeating module in the LSTM.
In (1), (2), (3), (4), (5), (6), (7) and (8), the forget gate $f_t$ determines how much the value of $c_{t-1}$ should be passed to $c_t$. The input gate $i_t$ scales the value of $g_t$ to compute $c_t$. $g_t$ is the input modulation gate. The output gate $o_t$ scales the value of the next hidden state $h_t$. $b_f$, $b_i$, $b_g$ and $b_o$ are the biases of the forget gate, input gate, input modulation gate and output gate respectively.

### 1.3.1.2 Gated Recurrent Unit (GRU)

Gated recurrent unit was introduced by Cho et al. [14] in 2014. The structure of gated recurrent unit is similar to LSTM. GRU cell consists of two gates which are a reset gate and an update gate. In some researches, GRU provides a comparable result to LSTM but for taxi demand prediction, it hasn’t been explored yet. Therefore, we selected GRU
to be another predictive model in this study. Figure 1.6 shows the structure of the repeating module in the GRU.

\[ z_t = \sigma(W_z \cdot [x_t + h_{t-1}]) \quad (9) \]
\[ r_t = \sigma(W_r \cdot [x_t + h_{t-1}]) \quad (10) \]
\[ h_t = \tanh(W \cdot [x_t + r_t \cdot h_{t-1}]) \quad (11) \]
\[ \tilde{h}_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot h_t \quad (12) \]

1.3.2 eXtreme Gradient Boosting (XGBOOST)

Gradient boosting tree model was proposed by Friedman et al in 2001 [16]. The main idea of boosting is to add new models to the ensemble sequentially. It combines many weak predictive models into a strong one by considering the errors made by the existing predictive models. A loss function, a weak learner and an additive model (to add weak learners to minimize the loss function) are three main elements in gradient boosting algorithm. Extreme gradient boosting (XGBoost) [15] is a library optimized for boosting algorithm. The library provides a scalable, portable framework which was used in many data science competitions. Figure 1.7 shows an example of the ensemble of Gradient Boosting Tree.
1.4 Statement of problem

The demand for transportation in the future is really difficult to obtain. The traditional surveys such as online survey, paper survey can’t be a reliable source in this case because there are too many areas in the city. To manage the transportation better, we need to know the demand for transportation in various areas. Furthermore, we need to understand the demand for transportation in different time of a day. The demand for transportation data in all areas with different time of a day, is too hard for a human to understand it perfectly. Nowadays, Taxis are equipped with GPS sensors. It provides many necessary data such as pick-up location, drop-off location, average speed of a trip, duration of a trip. Moreover, people can go to any place in the city by using taxi service. Therefore, Taxi GPS data is a good source that can refer to the demand for transportation. With data preprocessing procedure, we can transform the taxi GPS dataset into a sequence of taxi demand in a particular area. We can apply machine learning with the transformed data to create a predictive model which forecast the demand for transportation in the future.

1.5 Purpose of study

The purpose of this study is to develop a forecasting model which can predict the taxi demand hourly well with all types of area function in Bangkok, Thailand. The suitable time-series models such as long short-term memory (LSTM), gated recurrent unit (GRU) and extreme gradient boosting (XGBOOST) for different types of area function functions will also be investigated. The development of time-series model will be conducted using the knowledge of taxi driver picking-up/dropping-off behaviors learned from the real-world taxi GPS trajectories in Bangkok and historical weather.
report. At the end of this study, the developed model should be able to provide an
accurate prediction for taxi demand in Bangkok, Thailand. Moreover, in the data
analysis section, we will extract the taxi demand patterns of different types of area
function for both incoming and outgoing trips.

1.6 Significance of study

The developed forecasting model can provide taxi demand information in the
future. The model can also be used for urban planning to improve the traffic flow of the
city. We use historical weather data to enhance the performance of the model. The time-
series models such as long short-term memory (LSTM), gated recurrent unit (GRU),
and extreme gradient boosting (XGBOOST) are used to produce a better prediction. To
the best of our knowledge, there has been little research conducted on Bangkok taxi
demand using big data analysis. This study will be the very first work on forecasting
the taxi demand using taxi GPS trajectories in Bangkok, Thailand. Hence, it can be used
to satisfy the need of such a predictive model in the reality. Furthermore, in the taxi
data analysis section, the extracted peak taxi demand information should be able to
provide an opportunity to create a new business in transportation. This hidden trace can
make us understand the behaviors of both taxi drivers and passengers.
1.7 Contributions

- We discovered that the current state of art model (LSTM) can not provide the best prediction in all types of area function.
- We investigated the performances of LSTM, GRU and XGBOOST for predicting taxi demand in 7 types of area function.
- We proposed an ensemble model which can provide a better result than standalone models.
- We developed a taxi interactive visualization dashboard which can be interacted with users easily. We can explore the basic information of a huge dataset faster by using the developed dashboard.
- We showed the highest taxi demand time in 7 types of area function for both incoming trips and outgoing trips.

1.8 Outline

The rest of this thesis organized as follows.

- **Chapter 1** introduces the current problems of taxi service, motivation and machine learning models.
- **Chapter 2** Literature review about taxi demand prediction.
- **Chapter 3** Describe about the dataset, data preprocessing and basic statistics of Bangkok taxi datasets.
- **Chapter 4** Methodology for taxi demand prediction
- **Chapter 5** Interactive taxi visualization dashboard, the interesting insights and taxi demand pattern in7 types of area function
- **Chapter 6** we discuss about the result and conclusion
Chapter 2  
Literature Review

Figure 2.1 Heatmap of taxi pickup in Bangkok

Rich information from taxi equipped with GPS sensors drives many researchers to study on how to use this digital footprint to discover some knowledge that can help us improving the transportation system. Hence, there are many existing proposed systems which use taxi GPS trajectories as their transportation sensors.

Chang et al. [1], proposed a model that predicts taxi demand distribution with respect to time, weather condition and location by using three clustering algorithms, DBSCAN, K-Means, AHC. The model was developed by using 5 taxi drivers’ data over 2 months in Taipei, Taiwan. Moreira-Matias, L., et al [2], presented a model for predicting the number of services that will happen at taxi stands using the time-varying Poisson model and the auto regression integrated moving average (ARIMA). The research used the dataset from 441 vehicles with 63 taxi stands in the city of Porto, Portugal. Yuan, J. et al [3] from Microsoft Research Asia, used a trajectory dataset generated by over 33,000 taxis of Beijing, China in a period of 3 months to create a model for finding the fastest driving direction. They represented a road segment which frequently traversed by taxis as a node (landmark) then they constructed a time...
dependent landmark graph. With the landmark graph, they used the two-stage routing algorithm to find the fastest route.

Moreover, there are studies focusing on anomaly detection. Zhang, D. et al [4] applied isolation forest method which uses “Few and Different” to develop anomaly detection in taxi trajectories. The proposed model can detect anomaly trajectories from normal trajectories with the same origin-destination pair with high accuracy. This model can detect the behavior of taxi that try to drive in another direction to increase the distance. Moreover, road network changing can be detected with this model. The dataset is collected from over 7,600 taxis in Hangzhou, China for one year. B. Li et al [5] developed a model that discovers effective passenger-finding strategies. It uses a collection of feature patterns represented by triplet (time, location, strategy). The strategies can be either waiting or hunting. The GPS data generated by 4,528 taxis was used for analysis, the top 600 taxis out of 200 taxis as the positive examples and the bottom 600 taxis as the negative examples. The indicator of a taxi’s performance is determined by the accumulated distance covered by the taxi when it is occupied in the selected day. Yuan et al. [6] presented T-finder model which is taxi recommendation system based on the knowledge which derived from both passengers’ mobility patterns and taxi drivers’ pickup and drop-off. The model aims to provide parking places which have high probability to get a passenger and the routes to these parking places. Li et al. [7] proposed an improved ARIMA based prediction method to forecast the select 100 hotspots by not only considering the nearest historical taxi data but also the repeated pattern of pick-up and drop-off quantity in hotspots. An adaptive watershed algorithm is also proposed to split rough hotspots into smaller hotspots.

Traffic flow prediction is also a popular research topic recently due to the growing need for real-time traffic flow information in intelligence transportation. Y. Lv et al [8] presented a deep learning approach which used the stacked autoencoder model (SAEs) as building blocks to represent traffic features. The recent greedy layer-wise unsupervised learning algorithm is used to increase the performance of training procedure.
The dataset used in for taxi demand prediction doesn’t only come from taxi trajectory data but also comes from a taxi-booking application [9]. N. Davis et al [9], proposed a taxi demand forecasting model using multi-level clustering approach and data from a leading taxi booking application in the city of Bengaluru, India. Geohash [10], a geocoding system that can encode a geographic location into strings, is also used to represent areas of the city. Each location of the city was divided to geohashes. They applied linear time-series forecasting models such as ARIMA, Seasonal Naïve with those geohashes. Moreover, they proposed a multi-level clustering technique which was implemented based on those geohashes to improve the accuracy of selected linear time-series model. They focused on only the top 20 percent of the geohashes which have high demand of taxi. Jun Xu et al [10], recently presented a real-time taxi demand prediction model using geohash library [10] to divide the entire New York City into around 6,500 areas. They applied a mixture density network (MDN) over long short-term memory recurrent neural network (LSTM RNN) with a data sequence of the number of taxi requests in each area. LSTM RNN is the state of art sequence learning model that widely used in natural language processing. In this study, they didn’t directly predict the number of taxi requests but they predicted the entire probability distribution of taxi demand instead of deterministically predicting the number of requests for each area. Mukai and Yoden [11], also applied neural networks to predict the taxi demand for the 25 regions of Tokyo. The raw taxi GPS also included the region label which represented each area. They divided the hours of a day into 6 intervals (4 hours). The weather data also was used as a Boolean input to neural network which equals to 1 for raining and 0 for not raining. For neural network design, they used 1 hidden layer which has 50 neurons. They reported that the area that has highest error rate is the area that has many transportations such as railway, subway because the taxi demand is so small and non-periodical. Moreover, they also mentioned that day of week is a strong factor to do taxi demand forecasting. Kai Zhao et al [12], used yellow taxicab and Uber taxi in New York to build a predictive model to forecast the demand of taxi. They compared Markov predictor with Neural Network predictor. The result showed that Markov predictor can achieve high accuracy in the area with high theoretical maximum predictability while neural network performed better with lower theoretical maximum predictability.
predictability. They also showed that a simpler Markov predictor can outperformed Neural Network predictor in some occasions.

Those existing taxi demand prediction works focused on predicting taxi demand in all areas by dividing the entire city into specific small areas but there is no work that considers the impact of the area function on taxi demand prediction. This study is a very first study on demand prediction by considering area function of the area. In another part, we extract the basic statistical information of incoming trips and outgoing trips in many types of area function. Furthermore, we analyze the taxi GPS data to extract some hidden interesting insights in taxi services.
Chapter 3
Bangkok taxi dataset and Data Preprocessing

3.1 Bangkok taxi dataset

The dataset using in this study is real-world data generated by taxis GPS in Bangkok, Thailand from January 1 to June 30. The data collection process was supported by Toyota Tsusho Electronics (Thailand) Co, Ltd. The sampling rate of this dataset is 5 second. For historical weather data in Bangkok, we use the data from openweathermap.org which has hourly sampling rate. There are many types of weather condition obtained from the website such as Thunderstorm, Rain, Clouds, Clear, and Frog. we reclassify the data into only clear and rainy. The detail of each field in datasets is shown in table 1 below.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMEI</td>
<td>the taxi unique ID</td>
</tr>
<tr>
<td>Timestamp</td>
<td>timestamp of the sample point</td>
</tr>
<tr>
<td>Latitude/ Longitude</td>
<td>GPS location of the sample point</td>
</tr>
<tr>
<td>Speed</td>
<td>Current driving speed of the taxi</td>
</tr>
<tr>
<td>Angle</td>
<td>Current driving direction of taxi</td>
</tr>
<tr>
<td>Meter</td>
<td>Indicator whether the taxi is occupied or not (the status is 0 when available, 1 when occupied)</td>
</tr>
</tbody>
</table>
3.2 Data Preprocessing

First, we have to find the pick-up point, the drop-off point, duration, distance and average speed of each trip. The GPS sensor in a taxicab is very sensitive. Therefore, the coordinate position may sometimes be invalid. We want to create a reliable model to predict the demand of taxi accurately, thus the invalid data should be removed. Then, we will match the pick-up point with a place which is an area that we want to predict the taxi demand. Lastly, we will calculate the aggregated taxi pick-up in all selected areas hourly. Figure 3.1 shows the overview of data processing.

![Diagram of data preprocessing]

**Figure 3.1** The overview of data preprocessing
3.2.1 Extracting Origin-Destination

A shift of meter indicates a pick-up/drop-off event. If the meter is changed from 0 to 1, we will mark that coordinate as the pick-up point. On the other hand, if the meter is changed from 1 to 0, the coordinate will be marked as the drop-off point. In this study, we will use only occupied trips and ignore vacant trips. The example of pick-up/drop-off extraction is shown in figure 3.2.

![Figure 3.2 Origin-Destination Extracting](image)

We also calculate the distance, duration, average speed of each trip in this process.

3.2.2 Filtering invalid trips

We will filter the occupied trips which have abnormal duration and average speed. After preliminary investigating the average duration and speed of normal occupied trips, the normal range of duration should be 100 to 6,000 seconds. The normal range of speed should be between 1 m/s to 30 m/s. The trips which are out of those ranges will be removed.

3.2.3 Matching the pick-up event to POI

In this study, we use Points of Interest (POI) to match pick-up event with an area. If the pick-up event is within a radius of 200 meters from a point of interest, we will consider that the pick-up event occurred at that POI. Figure 11 shows the pick-up matching process.

![Figure 3.3 Matching the pick-up event to POI](image)
3.2.4 Calculating the aggregated demand in every hour for each point of interest

The number of pick-up events represents the demand of taxi. We aggregated the number of pick-up events for all areas in every hour. Therefore, we have the taxi demand of each area in 24-time slots a day.

<table>
<thead>
<tr>
<th>Time slots</th>
<th>0 - 1</th>
<th>1 - 2</th>
<th>1 - 3</th>
<th>.....</th>
<th>23 - 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Pick-ups</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>.....</td>
<td>10</td>
</tr>
</tbody>
</table>

**Figure 3.4** Example of aggregated demand in 24 time slots

3.3 Basic statistics

We also calculate the duration, average speed and distance of each trip in the data preprocessing. In this section, we explore the basics components of Bangkok taxi dataset.

**Figure 3.5** Heatmap of taxi demand per hour in Bangkok city

Figure 3.5 shows heatmap taxi demand per hour in Bangkok city. We can see the abnormal low taxi demand from new year festival during 1, 2 and 3 January, 2017. Moreover, Songkran festival causes unusual low taxi demand during 12 – 17 April, 2016. The taxi demand at 20.00 to 4.00 on weekend is clearly higher than the weekday. The taxi demand at 3.00 to 4.00 is the lowest taxi demand time of a day.
Figure 3.6 Distance of Bangkok taxi trips

Figure 3.6 shows that most of Bangkok city trips are short. This trend can confirm us that many taxi drivers in the city of Bangkok prefer short trip to long trip.

Figure 3.7 Duration of Bangkok taxi trips

Even, figure 3.6 shows that most of the taxi trips in Bangkok are short trips but many taxi drivers need to speed over 30 minutes for a single trip as we can see from figure 3.7. This situation may be caused by the traffic. Bangkok, Thailand has one of the worst traffic flow in the world. Although, we like to travel in short distance. We need to spend much time on the traffic.
Figure 3.6 shows the average speed of taxi in Bangkok. Bangkok drivers spent an average 64.1 hours a year in traffic jam. As we expected, taxi cab can drive between 20 to 40 km/h.

Figure 3.9 shows that the highest taxi demand hour in Bangkok city is 20.00 and the lowest taxi demand hour is 3.00. We can conclude that 18.00 – 23.00 are the hours which taxi drivers can cruise more passengers.
Chapter 4
Methodology for taxi demand prediction

This chapter we describe about the defined types of area function that we use in this research. Then, we talk about how we split a dataset into training sets and validating set. Apart from taxi GPS data, we also use weather data which was obtained from openweathermap.org as another weather. The details of weather data and all features we use for training the models can be found in sections 4.2 and 4.3 respectively. We explain the selected predictive models in section 4.4. In section 4.5, we describe about eliminating the bias of weight of different scale of inputs using normalization equation. Then, we show the evaluation metric for taxi demand prediction in section 4.6. Information about parameter tuning for Recurrent neural networks and extreme gradient boosting can be found in sections 4.6 and 4.7 respectively. Lastly, we show the architecture of our ensemble model in the last section 4.8.

4.1 Types of area function

- Airport
- Department Store
- Hospital
- Subway
- Residential area
- Education area
- Tourist attraction area

We select 10 places for each type of area function except the airport because Bangkok has only two airports. We obtained latitude and longitude of places from Google place API.
4.2 Train/Test split

We have a dataset generated from over 5,000 taxis from January to April 2016. We use the first 3 months data as a training set and the last month data as a validating set.

Training data

| Jan | Feb | Mar | Apr |

Validating data

4.3 Weather data

The weather data used in this study was obtained from [https://openweathermap.org/](https://openweathermap.org/) which represents only the overall weather in all areas of Bangkok.

![Example of weather data from openweathermap.org](https://openweathermap.org/

4.4 Features

There are 6 main features for the predictive model. These features are selected based on previous studies [11], [12].

- Number of pick-up
- Hour of a day
- Day of week
- Day of month
- Weather
- National Holiday
We also applied rolling window technique with the number of pickups feature. Let \( n(t) \) be the number of pick-up in time step \( t \) and \( w \) be the window size that we want to look back, if we want to predict \( n(t) \), we will also use \( n(t-1), n(t-2), \ldots, n(t-w) \) as extra features. The window size \( w \) is 24 in this study.

<table>
<thead>
<tr>
<th>n(t-1)</th>
<th>n(t-2)</th>
<th>n(t-3)</th>
<th>......</th>
<th>n(t-23)</th>
<th>n(t-24)</th>
</tr>
</thead>
</table>

Figure 4.3 Rolling window

4.5 Selected Predictive Models

- Long short-term memory (LSTM)
  It is the current state of art for taxi demand prediction.

- Gated recurrent unit (GRU)
  It was proposed in 2014. GRU can outperform LSTM in some tasks.
  At present, no researcher has used GRU for taxi demand prediction.

- Extreme gradient boosting (XGBOOST)
  It is very famous in many data science competition. It is portable and very easy to implement. We also want to evaluate this model for taxi demand prediction.

4.6 Normalization

We normalized the data to eliminate the bias of weight from different inputs using the formulation in (13).

\[
X_{\text{normalized}} = \frac{X - \text{Min}}{\text{Max} - \text{Min}} \quad (13)
\]

\( X \) is an input. \( \text{Min} \) is the minimum value of inputs while \( \text{Max} \) is the maximum value of the inputs.
4.7 Evaluation Metric

We evaluate the performance of our predictive models with the widely used prediction error metrics: Symmetric Mean Absolute Error (sMAPE). The formulation of the prediction error metric is given as (5).

\[
sMAPE_i = \frac{1}{T} \sum_{t=1}^{T} \frac{|Y_{i,t} - \hat{Y}_{i,t}|}{\frac{1}{2} (Y_{i,t} + \hat{Y}_{i,t}) + C} \tag{14}
\]

\(Y_{i,t}\) is the real demand while \(\hat{Y}_{i,t}\) is the predicted demand occurred in area \(i\) at time \(t\). We also include constant \(C\) which equals one to prevent division by zero.

4.8 Parameter tuning for RNNs

4.8.1 Number of hidden neurons

To find a number of hidden neurons in the model, we conduct an experiment with 100, 150, 200, 250, 300, 350 neurons. In the experiment, we use ADAM as the optimizer. For activation function, we use the hyperbolic in the hidden layer and linear in the output layer.

4.8.2 Optimizer

We evaluate the performance of both Long Short-Term Memory and Gated Recurrent Unit with three optimizers which are Stochastic Gradient Descent, RMSProp and Adam. Optimizer testing is attempted with 150 hidden neurons. For activation function, we use the hyperbolic in the hidden layer and linear in the output layer.

4.8.2 Number of hidden layer

We need to choose the number of hidden layers in the network. Therefore, we conduct an experiment with 1, 2, 3 layers. This experiment is attempted with 150 hidden neurons for each layer. For activation function, we use the hyperbolic in the hidden layers and linear in the output layer.
4.9 Parameter tuning for XGBOOST

We use grid search to choose the set of parameters that provides the most accurate prediction. The selected parameters are shown as follows.

- Evaluation metric: MAE
- Learning rate: 0.1
- N estimators: 150
- Max depth: 2
- Max delta step: 1
- Min child weight: 11

4.10 Ensemble model

We observed that a predictive model that provides a better prediction than the others lately is likely to produce more accurate prediction in the next time step. We develop an ensemble model that follows the observed trend. A single predictive model can’t provide the most accurate prediction for all areas. In the reality, it may take too much time to choose a specific model for a particular place. Therefore, we need a model that can adjust itself to produce a prediction based on the best predictive model for that area.

We combine the predictions from those three models to find the best final prediction by assigning more weight to the model that achieved low error rate in the previous $w$ time steps. We use sMAPE (Symmetric mean absolute percentage error) to evaluate the performance of each model. Firstly, we find the average of sMAPE in previous $w$ time steps for all models.

Let $E_m = \{e_{m(t-w)}, \ldots, e_{m(t-3)}, e_{m(t-2)}, e_{m(t-1)}\}$ be the set of sMAPE values for model $m$ in the previous $w$ time steps. We calculate the average sMAPE of model $m$ in the previous $w$ time steps using (13).

$$E_m(\text{avg}) = \frac{1}{w} \sum_{i=1}^{w} e_{m(t-i)}$$

(13)
Now, we know the performance of LSTM, GRU and XGBOOST models in the previous \( w \) time steps. Then, we rank those models based on their average sMAPE in the previous \( w \) time steps.

Let \( P_f \) be the final prediction and \( P_{m^{1st}}, P_{m^{2nd}}, P_{m^{3rd}} \) be the predictions from the models those were ranked in ascending order. \( \mu, \beta, \alpha \) are the coefficients of the first rank model, the second rank model, and the third rank model respectively. Finally, the final prediction \( P_f \) can be computed by the weighted arithmetic mean in (14)

\[
P_f = \mu P_{m^{1st}} + \beta P_{m^{2nd}} + \alpha P_{m^{3rd}}
\] (14)

Figure 4.4 illustrates the architecture of our ensemble model.

![Figure 4.4](image)
Chapter 5

Interactive taxi visualization Dashboard, interesting insights and taxi demand pattern of 7 types of area function

5.1 Interactive taxi visualization dashboard

Dealing with a huge dataset, we need a tool that helps us understand hidden information easier. The hidden interesting insights can be extracted by using the right data exploration tool. We use the following tools to implement an interactive taxi visualization dashboard.

- Crossfilter.js is a JavaScript library for exploring large multivariate datasets in the browser. We use it for grouping, filtering our Bangkok taxi datasets.
- D3.js is a JavaScript library for visualizing huge dataset using web standards.
- DC.js is a JavaScript charting library with native Crossfilter.js support. Moreover, it leverages d3 to render charts in CSS-friendly SVG format.
- Leaflet.js is a JavaScript library for interactive maps.
- Queue.js is an asynchronous library for JavaScript.
- Underscore.js is a JavaScript that provides functional programming helpers.
- A python library called flask is used to run the server side.
We sorted taxi drivers from their average number of trips per day. We select the first 400 drivers out of 4000 taxi drivers as the top taxi drivers. The others are considered as ordinary drivers. We will use top driver information to compare the behavior of top drivers with ordinary drivers. Then, we will show what top driver do to get a larger number of passengers than other drivers.

5.1.1 Data preparing for visualization

Before visualizing the taxi data on the dashboard, we need to transform the taxi dataset format into JSON format because we need to use the format that HTML can interpret the data. Figure 5.2 shows an example of a taxi data in JSON format.
The interactive dashboard provides the information as follows.

- Number of pick-ups
- Distance of a trip
- Average speed of a trip
- Day of week
- Day of month
- Hour
- Origin district
- Destination district
- Origin POI
- Destination POI
- Top driver trips.

5.2 Interesting insights

We use the developed taxi visualization dashboard in 5.1 to discover some interesting insights about taxi services in Bangkok, Thailand. Firstly, we describe the comparison of taxi demand patterns between the top and ordinary taxi in 5.2.1. Then, we check whether taxi demand can imply anomaly events. Lastly, we explore the top 10 taxi demand districts in Bangkok areas.
5.2.1 Comparison of taxi demand patterns between top and ordinary taxi drivers

![Graph of taxi demand patterns]

**Figure 5.3** Comparison of taxi demand patterns between top and ordinary taxi drivers using the taxi visualization dashboard

From figure 5.3, the taxi demand patterns between the top and ordinary taxi drivers are clearly different. Top taxi drivers have many trips in the night. From this observation, we can infer that the most of taxi drivers who have high number of trips are night-shift drivers. Day-shift taxi drivers tend to have a smaller number of trips comparing to night-shift drivers. To confirm this information, we calculate the number of trips by hour in figure 5.4.
5.2.2 Can number of taxi trips imply an anomaly event?

To check whether the number of taxi trips can imply an anomaly event, we choose Impact Muang Thong Thani which is a commercial complex consisting of an arena and exhibition halls as a studied place. Thai people don’t have many choices of transportation to go to Muang Thong Thani. Therefore, Taxi service is preferable. Figure 5.5 shows the number of taxi trips to Muang Thong Thani on February 2016. We also use our developed taxi visualization dashboard to obtain figure 5.5.

**Figure 5.5** Number of taxi trips to Muang Thong Thani on February, 2016
We investigate the Impact Muang Thong Thani schedules on February, 2016. The schedules showed that there are big events on 6 – 7, 12 – 13, 24 – 28 on February. It is very apparent that the demand of taxi can imply some anomalies.

5.2.3 Top 10 taxi demand districts in Bangkok

**Figure 5.6** Top 10 taxi demand districts for different time periods in Bangkok
From figure 5.6, Chatuchak district has most taxi demand in every time period in a day. During 00.00 – 05.00 h., Vadhana becomes the second highest taxi demand district. We have investigated this unusual trend. We found that many popular nightclubs are located in Vadhana district. The full detail about taxi demand of 50 districts in Bangkok can be found in appendix A.

5.3 Taxi Demand Pattern of 7 types of area function

In this section, we illustrate the basic statistic information of taxi in airport areas. The peak hour of taxi demand of each type of area function is displayed by using many data visualization types. We include both incoming trips and outgoing trips to the selected places. We briefly analysis taxi demand pattern in airport areas. The demand patterns of the other areas can be found in the appendices A.1, A.2, A.3, A.4, A.5 and A.6. Moreover, we report the peak demand of taxi hours for all types of area function in chapter 6.

5.3.1 Airport

5.3.1.1 Airport outgoing trips

![Figure 5.7 Heatmap of taxi demand at airport areas](image-url)

**Figure 5.7 Heatmap of taxi demand at airport areas**
Figure 5.8 Taxi demand at airport areas on different day of week

Figure 5.9 Taxi demand pattern on weekday and weekend at airport areas by hour
Figure 5.10 Distance, duration and average speed of outgoing trips at airport areas

From figures 5.7, 5.8, 5.9 and 5.10, we can clearly see that most of the trips occurred in airport areas are long trips (over 20 kilometers). Taxi drivers can drive with high speed. The peak demand hours are around 21.00 - 14.00. The smallest demand hours are between 3.00 and 4.00. People who traveled by an airplane usually go straight to their home with their heavy baggage. Taxi service is one of the most convenient transportations.
5.3.1.2 Airport incoming trips

**Figure 5.11** Heatmap of taxi demand to airport areas

**Figure 5.12** Taxi demand to airport areas on different day of week
From figures 5.11, 5.12 and 5.13, it is very clear that 4.00 – 7.00 are the highest taxi demand hours to go to airports. The smallest taxi demand hours are between 23.00 and 2.00. The most of taxi trips to airport areas have long distance and take over half an hour.
Chapter 6  
Experiment Result  

There are four sections in this chapter. In the first section, we find the optimal network parameters for Long Short-Term Memory and Gated recurrent unit. We evaluate those models with a different number of hidden neurons. Then, we show the comparison between 3 optimization algorithms. Moreover, the result of the experiment to find the appropriate number of hidden layers is included in the first section. In the second section, we find the optimal parameter for the ensemble model. The third section, we present the final results of LSTM, GRU, XGBOOST and the ensemble model. The last section, we report about taxi demand trend in 7 types of area function from taxi trips data analysis.

6.1 Optimal network architecture for recurrent neural networks  

6.1.1 Number of hidden neurons  
Table 6.1: sMAPE result of LSTM and GRU with different number of hidden neurons

<table>
<thead>
<tr>
<th>Neurons</th>
<th>sMAPE : LSTM</th>
<th>sMAPE : GRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>24.22</td>
<td>24.38</td>
</tr>
<tr>
<td>150</td>
<td>24.12</td>
<td>24.23</td>
</tr>
<tr>
<td>200</td>
<td>24.15</td>
<td>24.33</td>
</tr>
<tr>
<td>250</td>
<td>24.20</td>
<td>24.27</td>
</tr>
<tr>
<td>300</td>
<td>24.22</td>
<td>24.30</td>
</tr>
<tr>
<td>350</td>
<td>24.35</td>
<td>24.52</td>
</tr>
</tbody>
</table>

As we increased the number of hidden neurons in the network, there is no significant improvement in term of accuracy. The more number of neurons in the network, the more computation time it requires to train the model. Therefore, we use 150 neurons for the final architectures of LSTM and GRU.
### 6.1.2 Optimization algorithms

**Table 6.2: sMAPE result of LSTM and GRU with different optimization algorithm**

<table>
<thead>
<tr>
<th>Neurons</th>
<th>sMAPE : LSTM</th>
<th>sMAPE : GRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAM</td>
<td>24.12</td>
<td>24.23</td>
</tr>
<tr>
<td>RMSPROP</td>
<td>50.72</td>
<td>46.43</td>
</tr>
<tr>
<td>SGD</td>
<td>24.91</td>
<td>26.10</td>
</tr>
</tbody>
</table>

From Table 6.2, it is very apparent that using Adam as optimization algorithm provides the most accurate prediction. SGD achieves a comparable result to Adam while RMSPROP produces the worst prediction from three optimization algorithms.

### 6.1.3 Number of hidden layers

**Table 6.3: sMAPE result of LSTM and GRU with different number of hidden layers**

<table>
<thead>
<tr>
<th>Layers</th>
<th>sMAPE : LSTM</th>
<th>sMAPE : GRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24.12</td>
<td>24.23</td>
</tr>
<tr>
<td>2</td>
<td>24.08</td>
<td>24.34</td>
</tr>
<tr>
<td>3</td>
<td>24.10</td>
<td>24.20</td>
</tr>
</tbody>
</table>

Increasing the number of hidden layers does not always improve the performance of the model. As it shows in Table 6.3, both LSTM and GRU don’t generate a clearly better prediction with 2 hidden layers or 3 hidden layers. Hence, we select only 1 hidden layer in the network for the final architectures of both LSTM and GRU.

### 6.1.4 Final architecture of LSTM and GRU

From 6.1.1, 6.1.2 and 6.1.3, we found the optimal network architecture for LSTM and GRU. Therefore, the optimal parameters will be used for the ensemble model.

- Number of hidden neurons: 150
- Number of hidden layer: 1
- Optimization algorithm: ADAM optimization
- Activation function: hyperbolic tangent for hidden layers, linear in output layer
6.2 Optimal parameters of the ensemble model

6.2.1 Coefficients and lookback window

We evaluate the ensemble model with two sets of coefficients and look back window.

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\beta$</th>
<th>$\alpha$</th>
<th>window</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
<td>1</td>
<td>23.83</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
<td>2</td>
<td>23.78</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
<td>3</td>
<td>23.72</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
<td>4</td>
<td>23.74</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
<td>5</td>
<td>23.81</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
<td>6</td>
<td>23.83</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
<td>7</td>
<td>23.80</td>
</tr>
</tbody>
</table>

Table 6.4: sMAPE result of the ensemble model with coefficients (0.7, 0.2, 0.1)

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\beta$</th>
<th>$\alpha$</th>
<th>window</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>1</td>
<td>23.98</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>2</td>
<td>24.02</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>3</td>
<td>24.08</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>4</td>
<td>23.96</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>5</td>
<td>24.99</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>6</td>
<td>24.03</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>7</td>
<td>24.01</td>
</tr>
</tbody>
</table>

Table 6.5: sMAPE result of the ensemble model with coefficients (0.9, 0.1, 0)

From table 6.4 and table 6.5, we can conclude that the ensemble model with $\mu = 0.7$, $\beta = 0.2$, $\alpha = 0.1$ and loop back window = 3 hours, produces the lowest error rate. Therefore, I will use this set of coefficients for the final architecture of the ensemble model.
6.2.2 Final architecture of the ensemble model

We calculate the error rates of the three predictive models which are LSTM, GRU and XGBOOST. Then, we rank those three models based on their performance in previous ($w = 3$) time steps. We assign $\mu = 0.7$, $\beta = 0.2$, $\alpha = 0.1$ as the coefficients of the first rank model, the second rank model and the third rank model respectively. Lastly, we calculate the prediction of the next time step using (12). The performance of the ensemble can be found in section 6.3.

Figure 6.1 Final architecture of the ensemble model
6.3 Performance of the final models

<table>
<thead>
<tr>
<th>Areas</th>
<th>Models</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lstm</td>
<td>gru</td>
<td>xgboost</td>
<td>ensemble</td>
</tr>
<tr>
<td>All places</td>
<td>24.12%</td>
<td>24.23%</td>
<td>25.71%</td>
<td>23.72%</td>
</tr>
</tbody>
</table>

Table 6.6: Average sMAPE of all models in all places

Table 6.6 shows the average sMAPE of LSTM, GRU, XGBOOST and the ensemble model in all places. LSTM provides the best accuracy from those three standalone models. GRU generates a comparable prediction to LSTM. Even, GRU was proposed later than LSTM but LSTM is still better taxi demand prediction than GRU. Although, XGBOOST produces the worst prediction from those three standalone models, it is portable and easy to implement. It doesn’t require a good graphics card to produce a prediction quickly but both LSTM and GRU still need a good graphics card to train model quickly. Therefore, XGBOOST still is a good model for taxi demand prediction in some situation. The ensemble model can outperform all three standalone models with sMAPE of 23.72%.
6.3.1 Performance of the final models in 7 types of area function

<table>
<thead>
<tr>
<th>Areas</th>
<th>Models</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lstm</td>
<td>gru</td>
<td>xgboost</td>
<td>ensemble</td>
</tr>
<tr>
<td>Airport</td>
<td>18.60%</td>
<td>18.82%</td>
<td>17.54%</td>
<td>17.70%</td>
</tr>
<tr>
<td>Residential Area</td>
<td>24.01%</td>
<td>24.27%</td>
<td>27.05%</td>
<td>24.26%</td>
</tr>
<tr>
<td>Department Store</td>
<td>24.53%</td>
<td>24.58%</td>
<td>24.38%</td>
<td>23.78%</td>
</tr>
<tr>
<td>Education Area</td>
<td>22.15%</td>
<td>22.30%</td>
<td>27.29%</td>
<td>22.26%</td>
</tr>
<tr>
<td>Hospital</td>
<td>24.19%</td>
<td>24.30%</td>
<td>27.10%</td>
<td>24.41%</td>
</tr>
<tr>
<td>Subway</td>
<td>25.88%</td>
<td>25.89%</td>
<td>25.32%</td>
<td>24.90%</td>
</tr>
<tr>
<td>Tourist Attraction</td>
<td>25.15%</td>
<td>25.17%</td>
<td>24.73%</td>
<td>23.91%</td>
</tr>
</tbody>
</table>

Table 6.7: Average sMAPE of all models in 7 types of area function

The average sMAPE of all models in each area types were shown in table 6.7. The XGBOOST approach can outperform both LSTM and GRU model in the airport, department store and tourist attraction area which are high taxi demand areas. The prediction of GRU is really closed to LSTM’s but the result shows that LSTM still provides better prediction than GRU for all types of area. Both LSTM and GRU work very well in low taxi demand areas such as residential area, education area and hospital area. For the ensemble model, even though, it can provide the lowest error rate in only tourist attraction and department store areas but it produced the results that are closed to the result of the best model in each type of area function.
6.3.2 Comparison of predicted demand and actual demand

Figure 6.2 One-week taxi demand prediction at Don Mueang Airport

Figure 6.3 One-week taxi demand prediction at Chatuchak Subway station
In figures 6.2, 6.3 and 6.4, the blue line represents the actual taxi demand. The taxi demand prediction from XGBOOST is the green line. The taxi demand predictions of LSTM and GRU are the pink line and the yellow line respectively. We use the black line to represent the prediction from the ensemble model.

Those 3 figures illustrate that the model that provides a better prediction lately is likely to produce more accurate prediction in the next time step. The ensemble model was created used on the observed knowledge. It assigns more weight to the model that provides accurate prediction lately. Therefore, the ensemble model can provide a better than other standalone models in overall. The three figures confirm us that taxi demand in Bangkok areas is predictable in both low taxi demand and high taxi demand areas and the selected predictive models can be used to forecast the taxi demand in the reality.
6.4 The peak, off-peak hour of incoming trips and outgoing trips in 7 types of area function

<table>
<thead>
<tr>
<th>Places</th>
<th>Outgoing trips</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Peak hours</td>
<td>Off peak hours</td>
<td>Peak hours</td>
<td>Off peak hours</td>
<td></td>
</tr>
<tr>
<td>Airport</td>
<td>20.00 - 21.00</td>
<td>03.00 – 05.00</td>
<td>04.00 – 06.00</td>
<td>00.00 – 02.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14.00 – 16.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residential area</td>
<td>18.00 – 22.00</td>
<td>01.00 – 06.00</td>
<td>18.00 – 23.00</td>
<td>02.00 – 05.00</td>
<td></td>
</tr>
<tr>
<td>Department store</td>
<td>12.00 – 15.00</td>
<td>01.00 – 07.00</td>
<td>10.00 – 15.00</td>
<td>00.00 – 06.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>19.00 – 21.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education area</td>
<td>12.00 – 13.00</td>
<td>00.00 – 06.00</td>
<td>08.00 – 09.00</td>
<td>00.00 – 05.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>12.00 – 13.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital</td>
<td>10.00 – 15.00</td>
<td>01.00 – 08.00</td>
<td>05.00 – 07.00</td>
<td>00.00 – 04.00</td>
<td></td>
</tr>
<tr>
<td>Subway</td>
<td>12.00 – 13.00</td>
<td>03.00 – 06.00</td>
<td>11.00 – 13.00</td>
<td>00.00 – 04.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>19.00 – 23.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tourist Attraction</td>
<td>18.00 – 21.00</td>
<td>01.00 – 10.00</td>
<td>10.00 – 13.00</td>
<td>23.00 – 06.00</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.8: The peak, off-peak hour of outgoing and incoming trips in 7 types of area function

From table 6.8, the off-peak hour of taxi demand between 7 types of area function are quite similar. Going to the airports in the morning, people who live in Bangkok don’t have many choices. The most of convenient public transportations to Don Mueang Airport starts at 07.00 h. Therefore, the peak hours of incoming trips to airports area are around 04.00 – 06.00. The peak hours of incoming trips to hospitals can confirm us that Thai people have to go to public hospitals as early as they can because there are many patients waiting for medical services. They won’t get those medical services if they arrive at the hospitals late.
Chapter 5
Conclusions

First, we show the basic statistical information of taxi demand around Bangkok areas. The basic statistical information helps us understand the behavior of taxi and passenger in overview. Then, we proposed an ensemble model based on long short-term memory network (LSTM), gated recurrent unit network (GRU) and extreme gradient boosting (XGBOOST) models to predict taxi demand in 7 types of area function in the city of Bangkok, Thailand. We use Point of Interests (POIs) to match the taxi demand with studied areas. We find the optimal network parameters for both LSTM and GRU by conducting experiments. Parameter tuning for XGBOOST has been done to archive the most accuracy. We observed that the predictive model which provides a better prediction lately is likely to produce more accurate prediction in the next time step. The ensemble model predicts taxi demand by following this trend. The results show that the ensemble model outperforms other standalone models with sMAPE of 23.72% in all areas. For the independent models, LSTM provides the most accurate prediction in low taxi demand areas such as residential area, hospital and education area. LSTM achieves better results than GRU in all types of area. The XGBOOST model gives better results than the others in high taxi demand areas such as department store, subway and airport areas. Although, XGBOOST produces the worst prediction from those three standalone models in overall but it is portable and easy to implement. It doesn’t require a good graphic interface card to produce a prediction with reasonable computation time but both LSTM and GRU need a good graphic card to train model with workable speed. Therefore, XGBOOST still is a usable model for taxi demand prediction. This study shows that a single predictive model cannot provide the best prediction in all areas. Combining the predictions from various models can improve the performance in overall.

In the data analysis part, a number of taxi trips can also reflect the mobility of the city. The report of the peak and off-peak hour shows the behaviors of passengers. It shows the lacking of other transportations to specific areas at a period of a day. Moreover, it can an evidence to confirm social problems as we describe the peak hour of incoming trips to hospitals.
References


Appendix A
Additional data visualization

A.1 Department store

A.1.1 Outgoing trips from department store areas

Figure A.1.1: Heatmap of taxi demand at department store areas

Figure A.1.2: Taxi demand at department store areas on different day of week
Figure A.1.1.3 Taxi demand pattern on weekday and weekend at department store areas by hour

Figure A.1.1.4 Distance, duration and average speed of outgoing trips from department store areas
A.1.2 Incoming trips to department store areas

Figure A.1.2.1 Heatmap of taxi demand to department store areas

Figure A.1.2.2 Taxi demand to department store areas on different day of week
Figure A.1.2.3 Taxi demand pattern of department store areas on weekday and weekend by hour

Figure A.1.2.4 Distance, duration and average speed of trips to department store areas

54
A.2 Hospital

A.2.1 Outgoing trips from hospital areas

Figure A.2.1.1 Heatmap of taxi demand at hospital areas

Figure A.2.1.2 Taxi demand at hospital areas on different day of week
Figure A.2.1.3 Taxi demand pattern on weekday and weekend at hospital areas by hour

Figure A.2.1.4 Distance, duration and average speed of outgoing trips from hospital areas
A.2.2 Incoming trips to hospital areas

Figure A.2.2.1 Heatmap of taxi demand to hospital areas

Figure A.2.2.2 Taxi demand to hospital areas on different day of week
Figure A.2.2.3 Taxi demand pattern of hospital areas on weekday and weekend

Figure A.2.2.4 Distance, duration and average speed of trips to hospital areas
A.3 Subway

A.3.1 Outgoing trips to subway areas

Figure A.3.1.1 Heatmap of taxi demand at subway areas

Figure A.3.1.2 Taxi demand at subway areas on different day of week
Figure A.3.1.3 Taxi demand pattern on weekday and weekend at subway areas by hour

Figure A.3.1.4 Distance, duration and average speed of outgoing trips at subway areas
A.3.2 Incoming trips to subway areas

Figure A.3.2.1 Heatmap of taxi demand to subway areas

Figure A.3.2.2 Taxi demand to subway areas on different day of week
Figure A.3.2.3 Taxi demand pattern of subway areas on weekday and weekend by hour

Figure A.3.2.4 Distance, duration and average speed of trips to subway areas

Ref. code: 25605922040588APL
A.4 Residential area

A.4.1 Outgoing trips from residential areas

Figure A.4.1.1 Heatmap of taxi demand at residential areas

Figure A.4.1.2 Taxi demand at residential areas on different day of week
Figure A.4.1.3 Taxi demand pattern on weekday and weekend at residential areas by hour

Figure A.4.1.4 Distance, duration and average speed of outgoing trips from residential areas
A.4.2 Residential area incoming trips

Figure A.4.2.1 Heatmap of taxi demand to residential areas

Figure A.4.2.2 Taxi demand to residential areas on different day of week
Figure A.4.2.3 Taxi demand pattern of residential areas on weekday and weekend by hour

Figure A.4.2.4 Distance, duration and average speed of trips to residential areas
A.5 Education area

A.5.1 Outgoing trips to education areas

Figure A.5.1.1 Heatmap of taxi demand at education areas

Figure A.5.1.2 Taxi demand at education areas on different day of week
Figure A.5.1.3 Taxi demand pattern on weekday and weekend at education areas by hour

Figure A.5.1.4 Distance, duration and average speed of outgoing trips from education areas
A.5.2 Incoming trips to education areas

**Figure A.5.2.1** Heatmap of taxi demand to education areas

**Figure A.5.2.2** Taxi demand to education areas on different day of week
Figure A.5.2.3 Taxi demand pattern of education areas on weekday and weekend by hour

Figure A.5.2.4 Distance, duration and average speed of trips to education areas

70
A.6 Tourist attraction

A.6.1 Outgoing trips from tourist attraction areas

Figure A.6.1.1 Heatmap of taxi demand at tourist attraction areas

Figure A.6.1.2 Taxi demand at tourist attraction areas on different day of week
Figure A.6.1.3 Taxi demand pattern on weekday and weekend at tourist attraction areas by hour

Figure A.6.1.4 Distance, duration and average speed of outgoing trips from tourist attraction areas
A.6.2 Incoming trips to tourist attraction areas

Figure A.6.2.1 Heatmap of taxi demand to tourist attraction areas

Figure A.6.2.2 Taxi demand to tourist attraction areas on different day of week
Figure A.6.2.3 Taxi demand pattern of tourist attraction areas on weekday and weekend by hour

Figure A.6.2.4 Distance, duration and average speed of trips to tourist attraction areas
A.7 Taxi demand of 50 districts in Bangkok

Figure A.7 Taxi demand of 50 districts in Bangkok