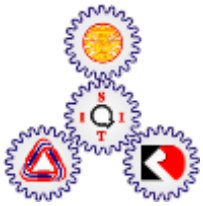

HOW DOES TAXI DRIVER BEHAVIOR IMPACT THEIR PROFITS? - DISCERNING THE REAL DRIVING FROM LARGE SCALE GPS TRACES

BY

THANANUT PHIBOONBANAKIT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE (ENGINEERING AND TECHNOLOGY)

SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY

THAMMASAT UNIVERSITY

ACADEMIC YEAR 2016

HOW DOES TAXI DRIVER BEHAVIOR IMPACT THEIR PROFITS? - DISCERNING THE REAL DRIVING FROM LARGE SCALE GPS TRACES

BY

THANANUT PHIBOONBANAKIT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE (ENGINEERING AND TECHNOLOGY)
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2016

HOW DOES TAXI DRIVER BEHAVIOR IMPACT THEIR PROFITS? - DISCERNING THE REAL DRIVING FROM LARGE SCALE GPS TRACES

A Thesis Presented

By

THANANUT PHIBOONBANAKIT

Submitted to

Sirindhorn International Institute of Technology

Thammasat University

In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE (ENGINEERING AND TECHNOLOGY)

Approved as to style and content by

Advisor and Chairperson of Thesis Committee _____

(Asst. Prof. Dr. Teerayut Horanont)

Committee Member and
Chairperson of Examination Committee _____

(Dr. Nguyen Duy Hung)

Committee Member _____

(Asst. Prof. Dr. Santi Phithakkitnukoon)

May 2017

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Asst. Prof. Dr. Teerayut Horanont for the continuous support of my master study and research, and for patient guidance, enthusiastic encouragement and useful guideline procedure during my study at GIS Lab, SIIT.

Besides my advisor, I am grateful to the chairperson of the examination committee, Asst. Prof. Dr. Boontawee Suntisrivaraporn and Dr. Nguyen Duy Hung, and a committee member, Asst. Prof. Dr. Santi Phithakkitnukoon, for their valuable advices and comments during my progress presentation.

Also, The Excellence Thai Student scholarship of Sirindhorn International Institute of Technology, Thammasat University, this support is the best opportunities in my life for study as graduated student.

This thesis would not be completed without kindness of our third-party company for supporting the mobility data and in the evaluation part of this thesis. Also, my senior and friends at the lab who have study and work in this field also deserve my sincerest thanks, their friendship and assistance has meant more to me than I could ever express. I could not complete my work without invaluable assistance of their participants.

Finally, my deepest gratitude also goes to my beloved family for their supports during my master study.

Abstract

HOW DOES TAXI DRIVER BEHAVIOR IMPACT THEIR PROFITS? - DISCERNING THE REAL DRIVING FROM LARGE SCALE GPS TRACES

by

THANANUT PHIBOONBANAKIT

Bachelor of Science in Information Technology (Second class honor), Sirindhorn International Institute of Technology Thammasat University, 2014

Master of Science (Engineering and Technology), Sirindhorn International Institute of Technology Thammasat University, 2017

With a trend towards the use of large scale vehicle probe data, the entire urban scale analysis is become possible to suggest useful information for taxi drivers and passengers. In our research we used Rama, I as representative of Phra Na Korn Side and BTS Wong Wien Yai as the representative of Thonburi side to make comparison. We also introduce taxi trip assessment model to make evaluation on taxi trips

This study, first, we have data exploration process to find trend and pattern of our obtained data. Then we make use the advantage of geospatial tools such intersection and buffering to divide and clustering it into areas. Then we calculate profit using taxi trajectory calculation algorithm by obtained fare rate and reconstructed trips. This process is running on the most advance tool of computing for the large-scale data such as Hadoop.

Second, the data were analyzed by using mathematic model to understand distance profit, total profit, total net profit in timely basis. The mathematic model is calculated taxi consumption cost, net profit, and energy resource cost. The taxi consumption cost is the cost occur from the normal drive and the consumption of the engine. The net profit is the cost where the total cost has been deducting by the consumption cost.

Third, we calculate probability of taxicab to observes chance that taxi drivers will get customer in the specific area. This could recommend taxi driver to make decision and plan the situation if they willing to come to this area or not.

Forth, we analyze on the taxi driver working hour and expense when they exit from the gas station. This would give the ground truth of actual cost that taxi driver will expense when they enter the gas station each day.

Finally, we built a model for fit and evaluate a result from our calculation model with real taxi trip which we collect from real driving of taxi drivers (Ground truth). We build to models. One is evaluate the accuracy of predict net profit and other model used to evaluate the classification of the worthiness level on the taxi trips. This step will be the final step to evaluate our work so far if it is conducting the correct way or not.

The result indicated that the pickup rate of taxi in this area is usually peak per area of interest and operation hours. The increased profits were mainly based on the distance and time of each trip. The trip run in shorter distance more frequency (9 times compare to one long distance) give high profit than the long distance. We also discover that most of taxi get customer from detour (On driving) in Rama I area. The result is varied on environment and area that we selected. Forth, taxi driver working hour is at least 14 hours per day from the data which we have been collated.
Finally, Random Forest regression is the easiest for model tuning and configuration over Random Forest and Decision Tree. It gives the prediction error only 10 Baht error from the actual net profit. We used the average profit from our ground truth it obtained over 90 percent accuracy. For the worthiness level classification, it took about 89 percent accuracy. In the evaluation process, we used Root Mean Squared and R-Squared to evaluated profit prediction model. Also, the worthiness classification model, we used confusion matrix to evaluate our model. These results uncover taxi driving behavior in Bangkok and yields great benefit for both taxi drivers and passengers. We also come up with solution to make taxi driver earn more profit and reduce decline service to customer by adding fare-rate to regular fare to make drivers earn more profit. In distance, less than 10 kilometers will add fare-rate 15 baht, 10 to 20 kilometers will add 30 baht and more than 20 kilometers will add 45 baht.
Keywords: Recommendation System, Taxi Profit, Data Mining

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Introduction and Theoretical Framework

Now a day, cost of living in Thailand has increase frequently. This outcome has been impact to many occupations especially on taxi driver. At this point, taxi driver has face many problems concurrently with their living. Some of them got an income less than the outcome or personal expense per day so they try to seek for opportunities to earn more income by a legal and non-legal way. For the customer side also face with problem of taxi meter decline to conduct the service from the following outcome above.

This is the main problem which we have face from the past until present day. Since we did not have a suitable solution to solve this issue.
To handle this problem, we conduct a taxi trip assessment model to evaluated trip from large scale GPS data. We extract taxi GPS point and combine it into trips. Then we input into algorithm to compute taxi cost, net profit and expense cost. From this result, we could recommend driver on place and time to create a trip to increase their profit. After we have a result then we create an evaluated model by apply regression technique such as random forest, gradient boost regression tree and decision tree. The model could be used to predict and forecast value of the trip which we input. The result is the trip value status that tell drivers if the destination is recommend to go or should be reconsider.

1.2 Statement of the Problem

To recognize taxi driver habit, behavior and actual profit in daily, we need a procedure that is efficient to classify and analyze input variable that collect from the mobility data. In addition, the procedure must not degrade the overall performance of the analyzing and predicting process. However, the way that our society use for collecting a data is to conduct a survey. For example, Paper Based survey, SMS Voting or Online Survey. This is the traditional way to get the input data from different location throughout our country. The information such as personal income and working log is very sensitive. Some of them ignore to give the true information so doing survey did not mean that it always gives direct accurate result. Some survey includes bias and variant input that did not related to a real-world situation.

If we still use this old technique to collect the data, it will waste time and money. In the case that the survey is not confidence or no one trust this result so we need to do the survey repeatedly in a cycle loop. It also causes a lack in the development stage if

we did not get the actual result on time.

A new technique to handle this issue is to use participatory system, which, apply machine-learning method. It can be mining and categorizing the data. The data is collecting from a large-scale GPS data, which get data from taxicab mobility device GPS periodically. For example, if you know taxicab trajectory pattern then you could make analysis compute income per trip, time usage, and actual net profit. This result could lead to recommendation on how taxi drivers can improve their profit and which time they should make a trip. In the future, we can use this information for taxicab recommendation system to forecast future trip profitable status.

1.3 Purpose of Study

The purpose of this study is to develop a recommendation system from large-scale taxi GPS probe data to provide the benefits and opportunities on Taxi service to the society. In this study, we focus on taxi trip from starting point to the destination. Try to come up with the result that which destination is give high profit to taxi driver. Also, give useful information which could improve our taxi transportation to be even better than the past.

1.4 Significance of the Study

Today many researches has introduced a new way to improve taxi trip profit in various method and framework. Most of them focus the recommendation on pick up place to have high chance to earn more profit. From the reviewing, some literatures, we discover that the remain question is that is the actual destination is valuable to route to and give high profit. Also, there are many unsolved questions that society need to have solution since taxi issue is popular in many urban capital areas. The popular on is the reason why taxi driver decline to conduct service to some customer.

In our work, we will introduce an algorithm to compute taxi trajectory cost, net profit, and time usage Also, trip valuable status from the predictive model which we use gradient boost regression tree to predict and evaluate our data. In this study, we could learn taxi movement pattern in each day. From this result, we desire to use the provide result for taxi trip assessment forecast system, which include taxi trip cost, net profit and trip profitable status.

Chapter 2

Literature Review

2.1 Previous Work

At the beginning of this research, we have studied on many literature from many researchers in mobility data field. Today many mobility devices had been introducing to our society since the first iPhone, which has been launched in 2007. From this event, it makes many changes to our society, habit, and interaction in our daily lives. People could use their device to serve there need, to interact with them as a personal assistant. From this point of view, we can use this information to predict activity and most significant visited place of individual within each day. In this field, many research had introducing new method. Also, using "Machine learning" method this study of data analysis and prediction. For example, use of location service and cellular network cites of mobile phone to collect most important places of individual (Scellato *et al.,* 2011). Make use of mobile phone sensor to predict and analyze motion and a way that individual interaction in their daily life and the outcome will visualize each transportation mode that individual select within a day (BYON and LIANG, 2014). Make use of Geographic information System technique such as create buffering and spatial data for analyze and predict selection of transportation mode that individual has been selected in each day of travel (Witayankurn *et al.,* 2013). Finally use route choice algorithm to identify the route path when individual select origin to their destination. First, Scellato, S., make use information from the location service and WIFI network log to determine an individual important location such as home and work places. In addition, the next most visited places (Scellato *et al.,* 2011). From this objective, a study conduct from location collection from each sample daily which concert about space location and time. It will divide data into two terms, which is data from GPS device that include in the smartphone devices, and data from WIFI data log. After data, has been collected, then will use machine learning method call "clustering" technique to classify data into cluster. From this point, similar characteristic or data set will group into the same cluster. This technique could be handle with a non-variant or certain data set. From this data category, could identify by MAC address when individual had connected to certain WIFI hotspot and GPS location data. For furthermore, they also apply spatio-temporal Markov predictor for predict the next place that individual is going to travel to from time history record which can use Markov theory to analyze data from the differentiation of time from the current and the last time that individual had been visited. This could estimate the next

visit that individual will visit this place again in the future. Also, Trinh Minh Tri Do, use mobile data to predict next visited and application from the past usage data log on the smartphone (Doa and D. G.-P., 2013). This could determine individual behavior where they frequently visit on each location. Also, could determine which application they are going to use in each time. Vincent Etter, introduce an algorithm call "Home change detection algorithm" to predict the stay and next placed visit and test with various type of machine learning such as dynamic Bayesian network, Artificial Neural Network, and Gradient Boosted Decision Trees (GBDTs). He discovers that using dynamic Bayesian network could predict more accurate result and use less performance when compare to artificial Neural network (Etter and M. K., 2013).

Second, BYON, Y. J., identify individual transportation mode from GPS and mobile sensor (BYON and LIANG, 2014). To achieve the research goal is to make use of GPS data and mobile phone sensor such as accelerometer and magnetometer that stream from smartphone to enhance and to make a result more accurate. From this two data, they use Neural Network to make analysis in pattern recognition from speed, motion, accelerate, and proximity of distance. After analyzing, data is classifying into modes such as universal mode detection for analyze all transportation mode in the city and auto mode detection for analyze in the traffic monitor purpose. To classify the pattern or physical characteristic, it can be classifying by the mobility movement, velocity, acceleration, magnetic field, and satellite data. In the case of universal mode detection, it had discovered that Neural Network could be able to classify all transportation mode from a physical operational characteristic. However, when use the auto mode detection it increases the accuracy of a result as long as the learning process stay concurrently, the more accurate result is given. From this study, it gives an idea of how to analyze and predict individual transportation mode from a data collection of the smartphone, which give more accurate result than use a GPS data only.

Third, Witayankurn, A., introducing the way of using Geographical Information System or GIS to analyze the transportation data (Witayankurn *et al.*, 2013). The important procedure to applying this approach is that we need to recognize the significant place or place that individual spent most of their time in such location. They also introducing an algorithm call "Stay Point Attraction Algorithm" (Ashbrook and Starner, 2003)(Changqing Zhou *et al.*, 2007), which analyze most staying place of an individual and make use of cellular network and GPS data. In term of GPS data, the device collect location every 5 minutes to decrease a battery usage of a

smartphone device. Also, from the location of the cellular network sites. In this case, using cellular sites only there will be a problem when smart phone switches the connection from one site to a nearby site. This cause a location of the individual change event He or She did not move to anywhere. To solve this problem, the author introduces a method to act like a buffer, which create a boundary of individual stay place. After recognizing the stay point, now will create change point detection method for recognize pattern of movement such as velocity, time different, and location change. This will tell mode of transportation. Also, introducing spatial data to create a buffer to set up a boundary. For example, train line and road network. If the location of the individual drops down within the buffer area, then we can predict that the individual is using train or vehicles to travel. In the transportation mode classification, the author using Random Forest Algorithm of machine learning for pattern recognition stage. After, all data is classifying then now it can be visualizing on the Web-Based trip visualization.

Forth, Route Choice Model, in this field, route choice model is very useful when use to find a desire route of user, it also be a part of finding shortest path to a selected destination. E.J. Manley, introducing heuristic model to bounded the route from the selected origin and the destination of the urban area (Manley *et al.*, 2015). Tom Thomas, using route choice model to prove that orbital routes are more attractive as these routes avoid the busy city center (Thomas and Tutert, 2015). Which involve traffic condition to make contrast of this issue. Zhengbing He, use route choice model to identify travelers' route choice in one day and describe the diversity of route choice behavior, which he discovers that individual activity and characteristic contain diversity among other when selected the route for their destination (He *et al.*, 2014).

Fifth, Anastasios Noulas use analyze about 35 million check-ins made by about 1 million Foursquare users in over 5 million venues across the globe then analyze features capture information on transitions between types of places, mobility flows between venues, and spatio-temporal characteristics of user check-in patterns (Noulas *et al.*, 2012). Also, study combining all individual features in two supervised learning models such as linear regression and M5 model trees (Noulas *et al.*, 2012).

Finally, Miao Lin, review relevant results on uncovering mobility patterns from GPS datasets. Specially, it covers the results about inferring locations of significance for prediction of future moves, detecting modes of transport, mining trajectory patterns and recognizing location based activities (Lin and H., 2014). For the mobility data, is

still wide topic so we scope down the topic to study the behavior and habit of the taxi drivers to recognize their daily income and behavior. In this field, many researcher try to introduce a way to make recommendation system such as recommend on routing, pick up place, and place to wait for customer. The researchers tried to find the way to solve this issue. Hwang et al (Hwang *et al.*, 2015), propose a taxi recommended system for determining the next cruising location by using L-L graph model. Qu et al (Qu *et al.*, 2014), developed a cost-effective recommended system for taxi drivers. Kamimura et al (Kamimura *et al.*, 2013), present a recommendation system, called D-Taxi, which would inform taxi drivers where to find the next passenger using the latest picking-up/dropping-off. Ding et al (Ding *et al.*, 2013), define a new method called global-optimal trajectory retrieving (GOTR). Zhang et al (M. Zhang *et al.*, 2012), proposed a novel method of pick-up recommendation for taxi driver based on spatio-temporal clustering. Salanovaa et al (Salanova *et al.*, 2011), presented a review of the different models developed for the taxicab problems. Zheng et al (Zheng *et al.*, 2009), mines interesting locations and classical travel sequences in each geospatial region and GPS trajectory by using tree-based hierarchical graph (TBHG). Yuan et al (Yuan *et al.*, 2011), presented a recommendation for taxi drivers and people expecting to take a taxi, using the knowledge of i) passenger's mobility patterns and ii) taxi driver's pick-up behaviors learned from the GPS trajectories of taxicabs. Qi et al (Qi *et al.*, 2013), also presented a method to predict the waiting time for a passenger at a given time and spot from historical taxi trajectories. Yue et al (Yue *et al.*, 2009), used taxi trajectory data to discover attractive areas where people often visit. Zhang et al (D. Zhang and He, 2012), propose a cruising system, pCruise, for taxicab drivers to maximize their profits by finding the optimal route to pick up a passenger. Somkiadcharoen et al (Somkiadcharoen *et al.*, 2015), use taxi probe to discover protester area during year 2013 and 2014 in Bangkok. Finally, Phiboonbanakit et al (Phiboonbanakit and Horanont, 2016) had introduced an algorithm to calculate and compare taxi fare-rate from taxi GPS probe data. Zhang et al (Y. Zhang and Haghani, 2015) use gradient boost tree to predict and improve traveling time. Use taxi data to predict demand by Moreira-Matias et al (Moreira-Matias *et al.*, 2013). Prediction of number of taxicabs with wavelet neural network by Yingjun et al (Yingjun *et al.*, 2012). Prediction bus passenger demand from mobile usage by (Chunjie Zhou *et al.*, 2016). Design taxi routing and fare rate estimation mobile application by Bai et al (Bai and Wang, 2012). Marketing design for profit on demand transport service by Egan et al (Egan and Jakob, 2016).

In Bangkok, Thailand, taxicab has played an important role in public transportation.

However, many taxi drivers in Thailand have got any degree of financial problems, mostly, brought about by multifactor such as higher cost of living and energy costs, regression of Thai economic growth as well as the new incoming competitors of alternative transportation. This issue has been widely discussed by many sectors but no real solution has been reached.

In comparison to our work, we would use and adapt the technique from previous literature to our research which is a study of benefit of taxi benefit and income and how to improve it. The also discover the remain question which need to be solve is that most of the literature are try to introduce pick up place. But is it true that when they pick up from this location and the destination that they route are profitable. This is the important question which we need to solve. If we would come with the answer and solution, this information will be useful to our society.

2.2 Preliminary Work

At the beginning of this research, we have make a study on mobility data from taxi GPS data to determine cost and benefit which will be earn by the taxi driver and society. Also from this work, we could suggest whether increase the taxi fare-rate is reasonable? Who will earn benefit from this issue?

Our approach is divide into stages as follow: GPS data collection, Taxi Routing and Fare-Rate Calculation Algorithm, Consumer Survey, and Comparison of the data. The result from this approach suggest that increasing taxi fare-rate per distance is reasonable because the trend on consumer, they choose to route with taxi with the short trip distance route rather that long distance route like in the past. In another hand, taxi drivers will earn more benefit from the increasing rate if the consumer choose to route in the long-distance mode. In addition, consumer concern about the taxi quality and standard the most so we would suggest that if we increase the fare-rate, we need to improve the quality and standard of the service concurrently.

This study has shown that mobility data is very valuable. It could use to interpret in many field for identify the issue.

2.3 Remaining Question

How we going to recommend and prove that the destination which taxi driver has been request from customer is profitable to them from the large-scale GPS traces.

2.4 Restatement of Research Question

To recognize Taxi driver behavior and activity pattern, we need a procedure that is efficient to classify and analyze input variable that collect from the mobility data. In addition, the procedure must not degrade the overall performance of the analyzing and predicting process. Since the Taxi driver activity and behavior has more diversity and variety. The research question will be how we going to handle this behavior diversity, to give the suggestion on their driving behavior and possible destination route which is profitable to the taxi driver from the given current location to be more efficient than the previous literature.

2.5 Limitation and Delimitation

From this research problem, we could determine our limitation and delimitation from many factors. In our limitation, since the size of data is very large when determine the population from a big data source, which we collect from mobile GPS. In this case, we used apache Hadoop to run the computation job which reduce execution time and have efficient resources management that deal which large scale data. The data which we obtained also have some error so we need to filter the error data out of our dataset which is time consuming. In another hand, the delimitation of our research is that we going to use only one transportation mode that is taxi vehicle since the taxi issue is very famous in the urban capital area. The society are interested in how we going to return the solvable solution which never occur before.

Chapter 3

Methodology

3.1 Material and Method

In this study, we have study on Big Data analysis on Taxi GPS in Bangkok. The importance thing before make analysis we need to do data exploration and algorithm development to handle these data. In this stage, we have divide task in to four stages. For example:

- Data Preprocessing and Exploration
- Taxi Algorithm
- Data Analysis
    - Basic Statistic
    - Net profit, Expense Cost, and amount to be add to regular fare
    - Probability to get customer from assign area
    - Taxi Working Hour
- Taxi Profit Predictive model

3.2 Data Processing and Exploration

In the beginning, we have obtained taxi GPS data almost from five thousand taxi vehicles in Bangkok during January to May 2016. Per this data, it is the big data contain many information which may have valuable and error so we need to make exploration and clean the data outliner and error. First, we create a grid 1 by 1 kilometer for clustering data to each grid (assign area) and remove data which not located in the Bangkok as shown in Figure 3.1 and use grid 100 by 100 meter for clustering data to road segment network as Figure 3.2.

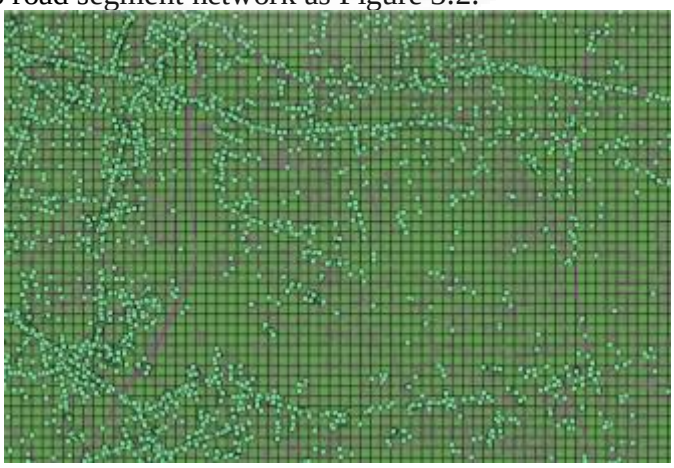

Figure 3.1: The taxi data are map to grid for area clustering

Figure 3.2: The road segment which we map to grid 100 meters

The example of the taxi trip transition from one state to another state which we need to analyze the valuable of taxicab trips and can give suggestion on trip routing are as shown in Figure 3.3. This could estimate the routing expense cost and the taxi net profit calculating from taxi probe movement.

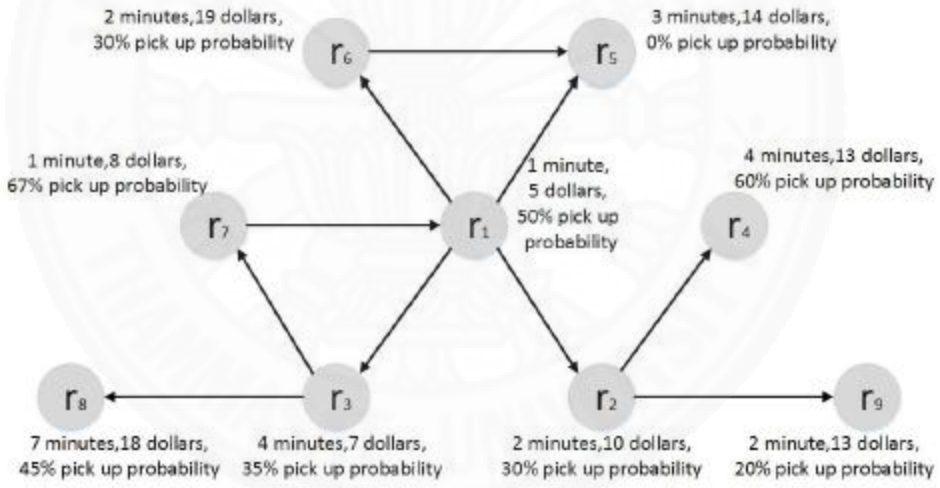

Figure 3.3: An example of route segment (Qu et al., 2014)

Second, we explore data component such as speed, data source, meter, date-time and IMEI. We remove unknown data sources because the data contain many data sources which we could not recognize. The usable data-source which we will use in this study are number 8 and 9. The data structure are as Table 3.1.

Table 3.1: Data attribute of taxi GPS probe

Field	Visualization	Notes
IMEI	10011304	Taxi Identification
Latitude	13.73522	Degree
Longitude	100.58979	Degree
speed	30	Move speed (km/h)
Direction	16	Degree
Error	0	Floating Point
Acceleration	0	Paddle press status
Meter	1	Status of taxi vehicle
Date-Time	2016-01-14 10:39:40	Time stamp of GPS point

3.3 Taxi Probe Data Extraction and Calculation

3.3.1 Fare-Rate Calculation Algorithm

We have developed an taxi fare-rate calculation algorithm (Phiboonbanakit and Horanont, 2016) to make taxi trip and calculate taxi vehicle trip distance, average speed, total trip time, cost, location of start and stop, and date time of start and end trip. We also obtained technique from "L-L Graph Model" (Hwang *et al.*, 2015) which extract importance factor of the data that transition from one cluster to the another cluster (In our study, we used term grid instead of cluster) such as distance, waiting time, trip time, traffic delay, and profit for analyzing impact to taxi driver's revenue. Each data point will calculate into origin and destination form. This algorithm is run on Apache Hive which is the useful tool to handle with the big data analysis.

3.3.2 Taxi Working Hour Detection

We developed an application to collect data of home, garage, and gas station. We would be able to recognize the working time of taxi drivers in Bangkok. Also, detect the location which is their home, garage or gas station.

3.4 Data Analysis

In this section, the data analysis is divide into 4 parts.
For example:

- Basic Statistic
- Net profit and Expense Cost
- Probability to get customer from assign area
- Taxi Working Hour

3.4.1 Basic Statistics

In this part, we explored the data statistic after we have generate a taxi trip from data point and try to figure out the actual trip is occurred in the reality and which one is occurred as an error. We explore in many components such as number of active taxi speed, distance, total trip time, traffic delay, and cost of each taxicab trips. Then finally we come up with a clean data that has remove all error according to the statistic description. In Figure 3.4, is an example of all taxi trip which we generate from taxi GPS probe. The data is categorized by date.

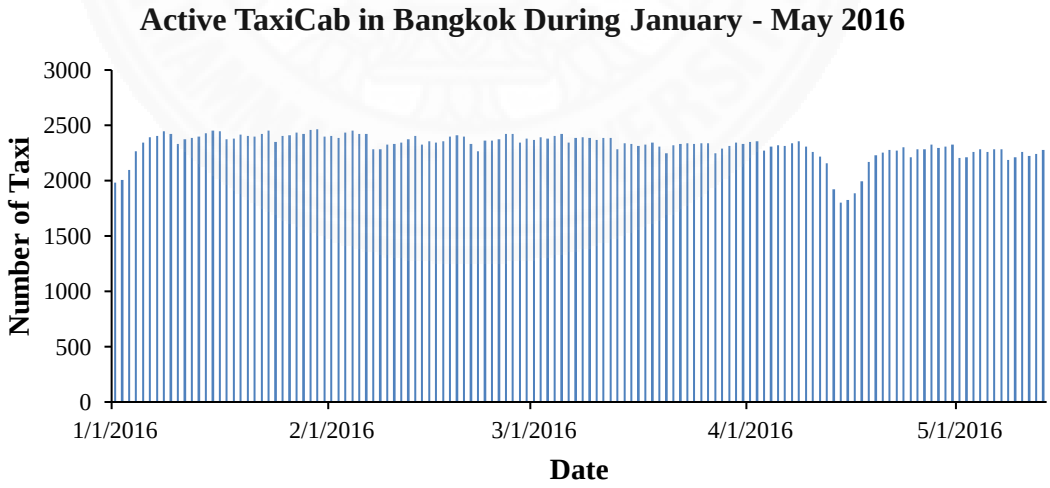

Figure 3.4: Active Taxicab during January to May 2016

Also, in Figure 3.5,Figure 3.6 ,Figure 3.7 ,and are the basic statistic of speed, distance, and trip time of taxicab trip.

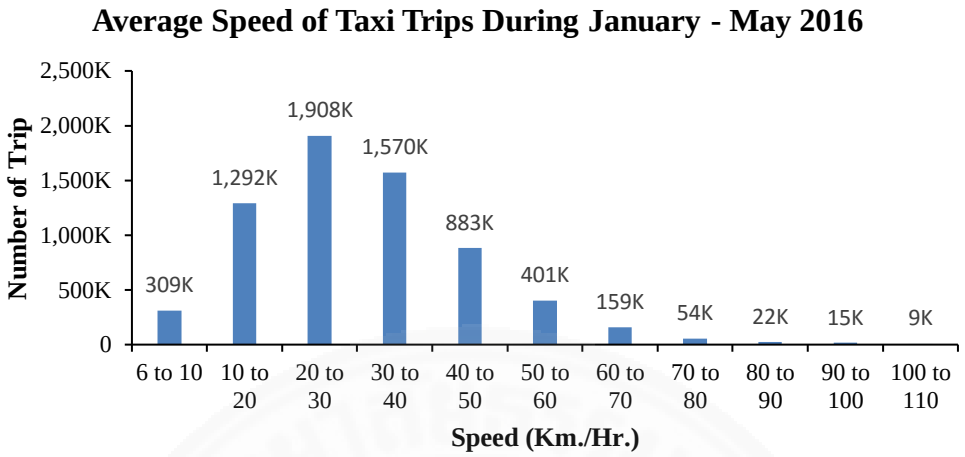

Figure 3.5: Average Speed of Taxicabs

From Figure 3.5, this could demonstrate that most of the taxicab trip in Bangkok has an average speed around 20 to 30 kilometers per hour due to traffic condition and some of the trip is the cruising of taxicab search for their customer.

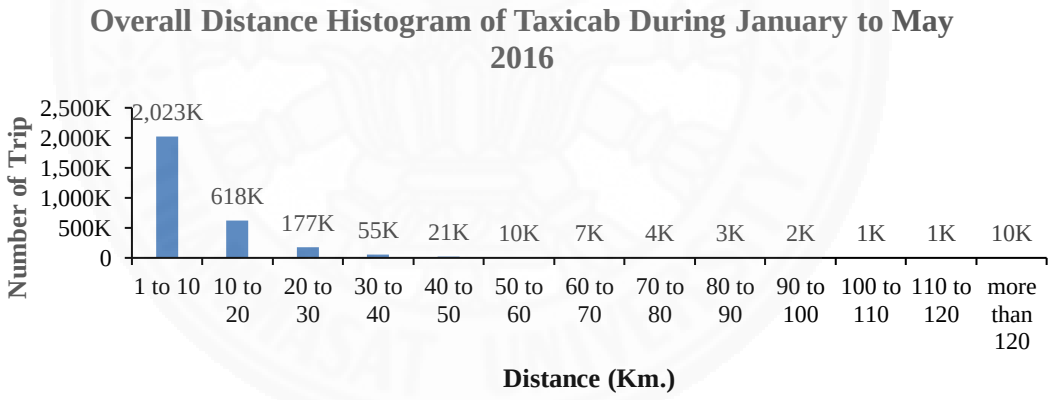

Figure 3.6: Distance of Taxicab's Trip

From Figure 3.6, We could demonstrate that most of taxi trip in Bangkok are most likely to be the short distance trip rather than running on the long run. This could point to a problem of taxi in Bangkok directly because some driver concert with these distances and think that short distance did not make any benefit to them. This topic also is our main case study. We also make validation with the rearrange the distance range data (Change from range every 10 kilometers to 5 kilometers) and the ground truth that we collect from real trip of 9 taxicabs as in appendix Figure B.13 and Figure B.14.

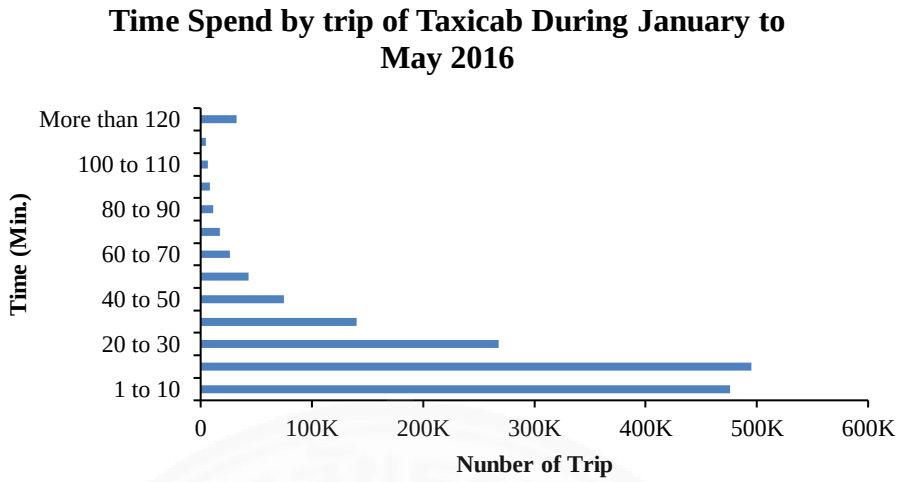

Figure 3.7: Time Spend of Taxicab's Trip

3.4.2 Net Profit and Expense Cost

In this part, we use the result from taxi fare rate calculation algorithm to calculate and find net profit and expense. The net-profit was calculated by extracting the service cost out of current profit as shown in the equation (1). This equation and input for calculation are described in section 3.

3.4.3 Probability to pick up customer from assigned area

we calculate the probability of taxi to get customer from assign area in each grid and group by hour. The formula to calculate the probability which taxicab will have customer onboard routing was shown in the equation (7). We estimate into two ways which is we set desire origin and unknown destination compare with we set desire destination and unknown origin to make comparison on place that impact to chance to get taxi to those locations.

3.4.4 Taxi Working Hours Inference given Home and Gas Station Locations

In this part, we used the result which obtained from mobile application for analysis stage. The result give the starting to work (Depart Home) and stop working (Arrive Home). We could have obtained drivers working hour. The application also give the location of taxi garage (if any) and famous gas station. We could know the actual cost when the driver pay for refill energy resources.

3.5 Taxi profit prediction model

For this part, we construct a model to make prediction and estimate accuracy of the built model. From algorithm 1, it provides many features to act as input to the model. Such as Source Direction, Destination Direction, Distance, location, Area grid id,

Traveling time, Speed, Max Speed, Min Speed, Traffic Delay, meter status, timestamp, and day of week. We apply Recursive Feature Elimination Method (Granitto *et al.*, 2006) which is wrapper method for feature selection. The candidate model which we have tested are random forest (Liaw and Wiener, 2002)(James *et al.*, 2015), decision tree (Nilsson, 2005), and gradient boosted regression tree (James *et al.*, 2015) in order to make profit prediction and estimate data accuracy. In finally, we used random forest as our predictive model since gradient boost regression tree have consume a lot of time computation. We build this model in Spark 2.0. For the model evaluation, we use Root mean square error(RMSE) and R-Squared to evaluate the prediction result of each model. The less RMSE is the efficient model of prediction. The feature is distance, speed, travel time, traffic delay, period of starting the trip, which we obtained from Recursive Feature Elimination Method.

3.6 Problem Definition

3.6.1 Road Network Formulation (Qu *et al.*, 2014)

Definition 1 (Road Segment). A long street can be separated into several road segments r by its crossroads. Specifically, each segment r is associated with a start point r.s and an end r.e. Moreover, each segment r also has several adjacent segments forming a set r.next[], which satisfies $\forall r_i \in$ r.next[] iff. r.e = r_i.s.

Definition 2 (Route). A route R is a sequence of connected road segments, i.e., R = ($r_1 \rightarrow r_2 \rightarrow \cdots \rightarrow r_M$), where $r_{k+1}.s = r_k.e$ ($1 \leq k < M$). The start point and the end of a route R can be represented as R.s = r_1.s and R.e = r_n.e.

Definition 3 (Road Segment Network). The road segment network G can be represented by a graph G =< V, E >, where V = {r_i} is the node set that consists of all road segments and E is the edge set, which satisfies $\exists e_{ij} \in$ E iff. $r_j \in r_i$.next[].

3.6.2 Paths and Connections

Connectedness of pairs of vertices in a graph G is an equivalence relation on V. Clearly, each vertex x is connected to itself by the trivial walk W := x; also, if x is connected to y by a walk W, then y is connected to x by the walk obtained on reversing the sequence W; finally, for any three vertices, x, y, and z of G, if xWy and yW′z are walks, the sequence xWyW′z, obtained by concatenating W and W′ at y,is a walk; thus, if x is connected to y and y is connected to z, then x is connected to z. The equivalence classes determined by this relation of connectedness are simply the vertex sets of the components of G (Bondy and Murty, 2008).

3.6.3 Fare-Rate Calculation Algorithm

The taxi fare calculation algorithm starts with determine the trip distance, travelling time, traffic delay then calculate the total fare-rate as Algorithm 3.*1*. The taxi-fare rate (Baht per distance) level is changed due to the distance range obtain from Department of Land Transport or DLT (Terry, 2014). For more detail of the function please see in Appendix A.4 Distance Calculation Function, A.5 Time Calculation Function, and A.6 Findcost calculation Function.

Algorithm 3.1 Taxi Fare-Rate Calculation Algorithm

```
Initialize cost, totaltriptime , traffic delay and distance to zero
Initialize Array for speed
Initialize p_meter, p_grid, p_lat,p_lon and p_dt to "None"
While data I less than total data size
    If p_meter is equal to "None" or "0" and meter is equal to meter is equal to "1"
        Set p_meter to meter,Set p_grid to grid, Set p_dt to dt
        Set p_lat to lat and olat to lat, Set p_lon to lon and olon to lon
    Else if p_meter is equal to "1" and meter is equal to meter is equal to "1"
    Set distance equal the sum of current distance and distance from distance calculation
function
    Set Total trip time equal the sum of current totaltriptime and time from findtime
function
    If speed less than 6
        Set traffic delay equal the sum of current trafficdelay and time from findtime
function
    Input speed to array Speed
    Set p_meter to meter,Set p_grid to grid,Set p_dt to dt,Set p_lat to lat
    Set p_lon to lon
    Else if p_meter is equal to "1" and meter is equal to meter is equal to "0"
    Set distance equal the sum of current distance and result from distance calculation
    function
    Set Total trip time equal the sum of current totaltriptime and result from findtime
function
    Set cost equal the input cost and traffic delay to findcost function
    Set dlat to lat and dlon to lon
    Print imei,lat,lon,olat,olon,dlat,dlon,distance,totaltriptime,traffic delay,cost,dt
```

3.6.4 Net-profit and Expense Cost Calculations

In this part, we use the result from our taxi fare-rate algorithm to calculate and find net profit and expense. The net-profit was calculated by extracting the service cost out of current profit as shown in the equation (1):

$$\text{NetProfit (N)} = \text{Cost} - \left(\frac{\text{d}*\text{ct}}{\text{md}}\right) - (\text{T} * \text{S}) - \text{TF} - \left(\frac{\text{pd}*\text{ct}}{\text{md}}\right) - (\text{PT} * \text{S}) - \text{PTF} \quad (1)$$

where d = distance by hour, T = total trip time, pd = distance before pick up customer and PT = total trip time before pick up customer.

In the present study, Toyota Corolla Altis 1.6 CNG, which commonly used as the taxi vehicle, was used as a sample vehicle which has 55 liters of fuel tank and 75 liters of NGV tank. The fuel consumption is 12.19 km/L obtained from manufacture eco sticker. The price is 22.04 Baht per liter for fuel and 13.36 Baht per liter for NGV (applied on May 5,2016). "ct" stands for cost as full tank. The calculation was shown in the equation (2):

$$\mathrm{ct} = \mathrm{fueltank} * \mathrm{fuelprice} \qquad (2)$$

md denoted as maximum distance which vehicle can drive from one full tank of fuel or NGV which can be calculated from the fuel consumption and total amount of fuel in one tank; equation (3)

$$\mathrm{md} = \mathrm{fuelconsumption} * \mathrm{fueltank} \qquad (3)$$

S denoted as service cost which calculated from taxi rental cost in Thailand which is about 1,000 Baht per day and divided by 24 Hours to find cost per hour and finally divided by 60 to find cost per minutes as shown in the equation (4.1):

$$\mathrm{S} = \frac{\frac{\text{taxi rental cost}}{24}}{60} \qquad (4.1)$$

TF denoted the consumption of vehicle when stop or run slow for a long time. To find vehicle consumption as shown in equation (4.2)

$$\mathrm{TF} = \left(\frac{\text{traffic delay} * 20}{1000}\right) * \text{fuel price} \qquad (4.2)$$

Note that PTF use the same calculation equation as TF where PTF is the calculation on traffic delay of trip before pick up customer but TF is for the current trip which pick up customer already.

To calculate Net profit per distance, we substitute "N" from equation (1) and "d" from the total distance by hour show in equation (5):

$$\mathrm{C} = \frac{\mathrm{N}}{\mathrm{d}} \qquad (5)$$

Then the average variable got from each trip obtained from equation (5) would be multiplied with total distance of each hour as shown in the equation (6) to get the profit:

$$\text{Profit} = \left(\frac{\mathrm{C1}+\mathrm{C2}+\mathrm{C3}\ldots+\mathrm{Cn}}{\mathrm{Cn}}\right) * \text{Total distance in each hour} \qquad (6)$$

3.6.5 Probability to Pick Up Customer Departing from Selected Area

After that, we calculate the probability of taxi to get customer from assign area in each grid and group by hour. The formula to calculate the probability which taxicab will have customer onboard routing was shown in the equation (7).

$$p = \frac{\Sigma(\text{Trip meter}=1)}{\Sigma(\text{all taxicap trip in that grid})} \quad (7)$$

3.6.6 Taxi Trip with Additional Fare Model to Solve Profit Loss

In this section, we create a mathematic model to create a return fare to make drivers more profit table. Also, calculate amount to be add in the trip to be satisfy to both customer and drivers as equation 8.

$$RT = \big((PL * \text{Distance}) * \text{Probability}\big) + (IP - \text{NetProfit}) \quad (8)$$

Where RT denoted Return Cost which we should add to regular fare-rate, PL denoted profit loss of each trip and IP denoted Ideal Profit which taxi hope to obtained.

Chapter 4

Taxi Trip Assessment

4.1 Taxi trip assessment model

In this section, we would describe out proposed model which consists of 5 stages as follow:

4.1.1 Stage 1: Determine taxi trip cost

In stage 1, we apply Algorithm 3.1 to our large-scale GPS probe to create a taxi origin and destination trip and it cost. The taxi trip is consisting of theses information as follow:

- IMEI, Distance, Trip time, Traffic Delay, Timestamp, Cost, Meter status, Origin grid and destination grid

4.1.2 Stage 2: Calculate net-profit and expense cost

In stage 2, we input the cost, distance, trip time and traffic delay which we obtained from the taxicab trip to calculate the actual net-profit which they have earned from the taxi routing. Also, the cost they have spent before and after done the trip. We apply procedure from section 3.6.4 to done the process.

4.1.3 Stage 3: Taxi Trip Value Estimation

After we determine the actual net-profit and expense cost, then we could determine the actual value of taxi trip by calculate the net profit by distance which has been describe in section 3.6.4. This could determine the level of valuable of taxi trip, the large number of net-profit per distance is give the high valuable on taxicab trip. We also create a model for solve and calculate additional fare-rate to be add.

4.1.4 Stage 4: Taxi working hour and pick up behavior detection

In addition, we make stage 4 to be supporter of stage 3 as we determine the behavior of taxi driver such as their working hour and their customer pick up behavior which could use to support the suggestion result of taxicab valuable.

4.1.5 Stage 5: Taxi Profit Model Evaluation

In this stage, we used all information which we have got to make suggestion on taxi trip and answer the following question as follow:

- How new taxi fare-rate make taxi-driver to gain more benefit?
- Which time that we will have more run and get more customer?
- Which type of routing that we need to choose to get more profit?
- How long that taxi driver work per day?

From the data, which we have obtained, we could create a prediction model to valid our data accuracy and make prediction on the future taxicab routing worthiness value. Also, determine which prediction model is suitable to our data that answer the following question:

- What is the future net-profit prediction model?

We Use Root Mean Squared Error(RMSE) to evaluated the error and how spread between observed and expected data in our profit model as formula shown in Figure 4.1.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

Figure 4.1: Root Mean Squared Error Formula

We also use R-Squared for ability to find the likelihood of future events falling within the predicted outcomes as formula shown in Figure 4.2 .

$$\mathbf{r} = \frac{\mathbf{n}(\Sigma \mathbf{xy}) - (\Sigma \mathbf{x})(\Sigma \mathbf{y})}{\sqrt{[\,\mathbf{n}\Sigma \mathbf{x}^2 - (\Sigma \mathbf{x})^2\,][\,\mathbf{n}\Sigma \mathbf{y}^2 - (\Sigma \mathbf{y})^2\,]}}$$

Figure 4.2: R-Squared Formula

Finally, we calculate Mean Absolute Error(Mae) to measure how close predictions are to the real values. The smaller the value for MAE, the better the algorithm in performance. We describe as follow equation (9):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y1 - y2|^2 \qquad (9)$$

For the worthiness level classification, we use confusion matrix to obtained true positive(TP) where it is label in this class and prediction is true, true negative (TN) where it is label in this class but classify as false, False positive (FP) where is label it is not in this class but in classify as true, and false negative (FN) where it label is not in this class and classify as false as in Table 4.1.

Table 4.1: Confusion Matrix

	Trip actual belongs to the category	Trip not belongs to the category
Classify Say the Trip Belong to the category	True Positive (TP)	False Positive (FP)
Classify Say the Trip not belong to the category	False Negative (FN)	True Negative (TN)

The formulas of calculation are as follow:

4.1.5.1 Accuracy

$$\text{Accuracy} = \frac{TP+TN}{Total} \tag{10}$$

4.1.5.2 Misclassification Rate

$$\text{Misclassification Rate} = \frac{FP+FN}{Total} \quad \text{or } 1 - \text{Accuracy} \tag{11}$$

4.1.5.3 True Positive Rate

$$\text{True Positive Rate} = \frac{TP}{Actual\ yes} \tag{12}$$

4.1.5.4 False Positive Rate

$$\text{False Positive Rate} = \frac{FP}{Actual\ No} \tag{13}$$

The overall Framework which include trip assessment model is shown in Figure 4.3.

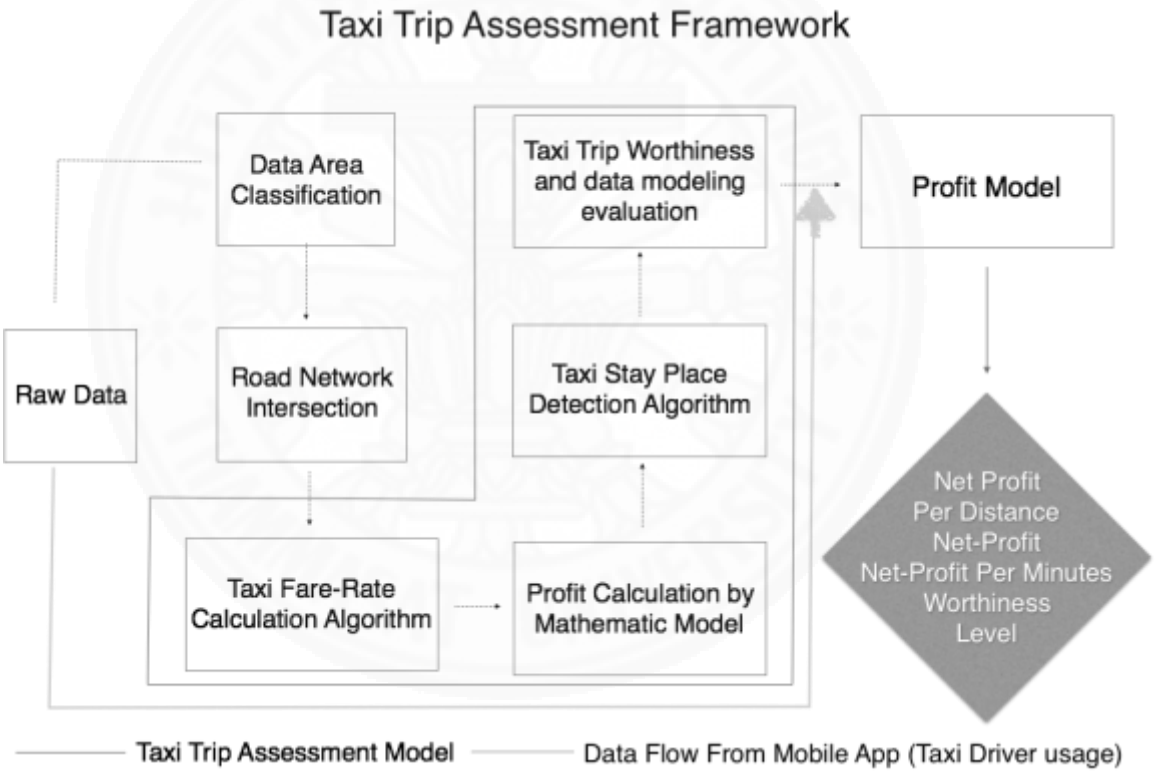

Figure 4.3: Trip Assessments Framework

Chapter 5

Result and Discussion

5.1 Results of Data Analysis and Prediction

Our study result will divide into 7 parts as follows:

- Taxi Survey
- How new taxi fare-rate make taxi-driver to gain more benefit?
- Which time that we will have more run and get more customer?
- Which type of routing that we need to choose to get more profit?
- How long that taxi driver work per day?
- What is the future net-profit prediction model?
- How much fee that we need to add to make drivers get more profit?

5.1.1 Paper-Based Taxi Survey

In this part, we used paper-based survey to support our study. We have different type of taxi in Bangkok such as public and private taxi as Figure 5.1 .

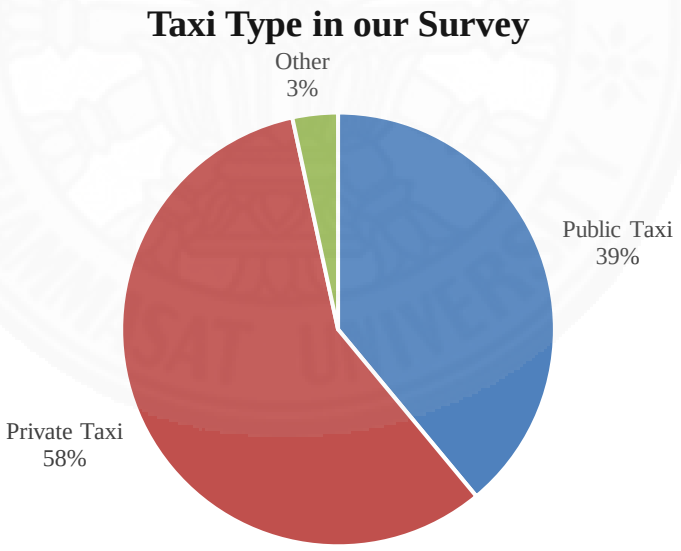

Figure 5.1: Type of Taxi in Our Survey

Then, they would like to cruise for customer most where economic and department store were located as Figure 5.2.

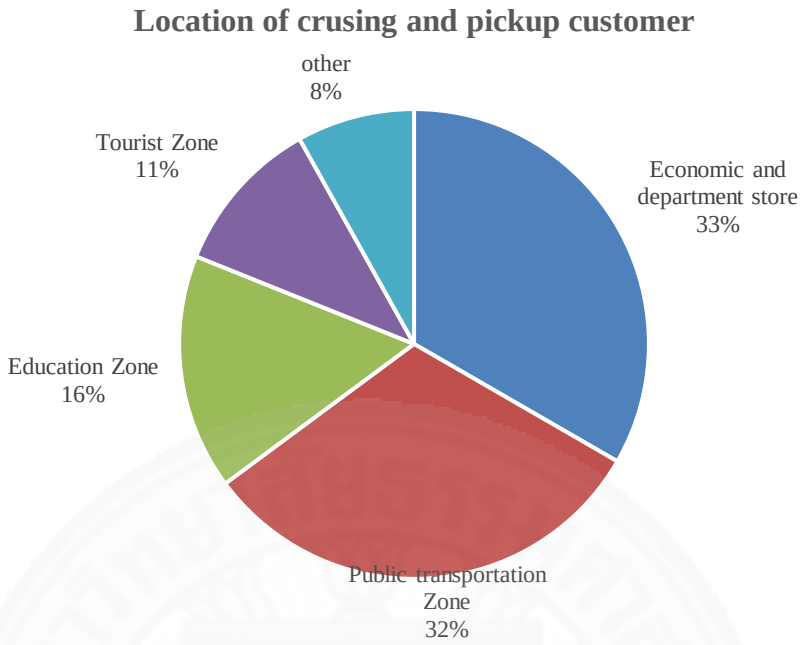

Figure 5.2: The location where taxi driver cruising and pick up for customer

Finally, the important information which our society need to discover is that the reason why taxi drivers decline conduct a service on customer. The survey tells that 62% of them has experience in decline service to customer. Also, most reason is the required routing is differed from prefer route as Figure 5.3 and Figure 5.4.

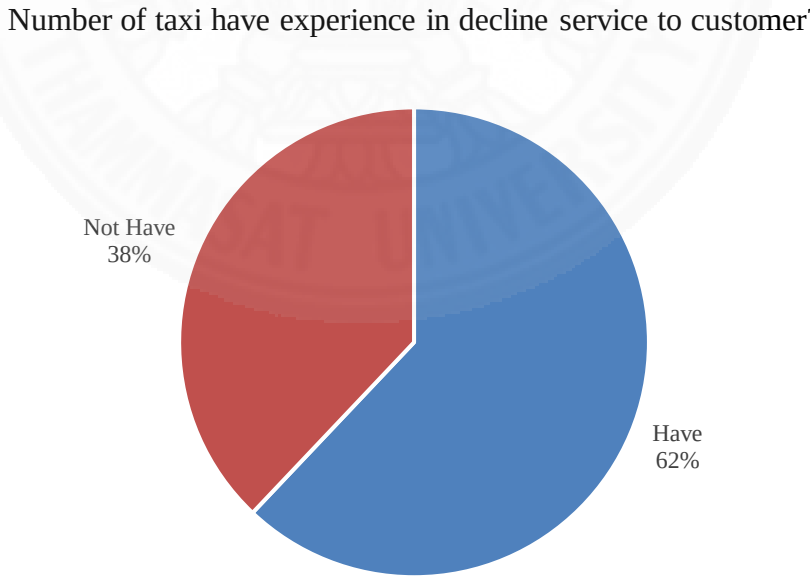

Figure 5.3: Have they ever decline service to customer?

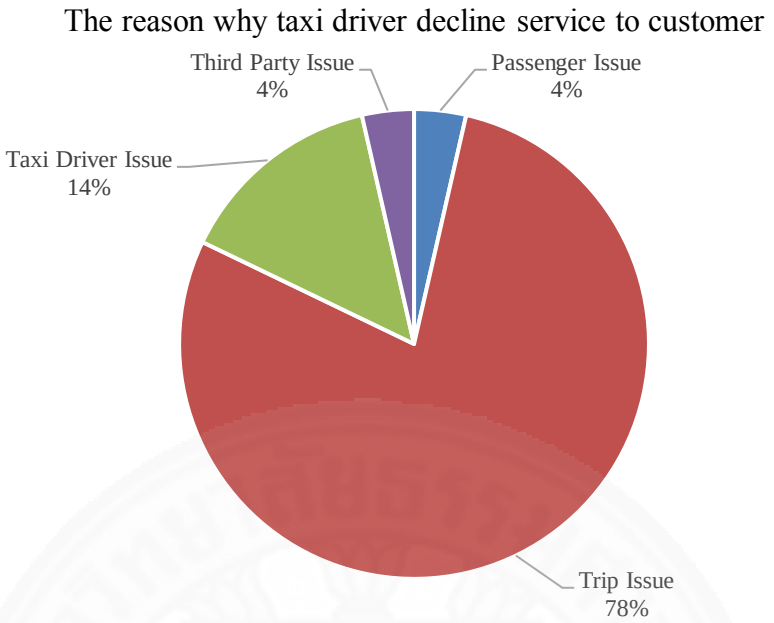

Figure 5.4: The reason of taxi driver why the decline service on customer

5.1.2 How does new taxi fare-rate gain taxi driver to gain more benefit

At the beginning, we develop an algorithm to calculate taxi fare-rate from taxi GPS point as Algorithm1. This algorithm will give a comparison of two version of fare-rate such as the fare-rate that obtained in Year 2014 and year 2015 (Phiboonbanakit and Horanont, 2016) as Table 5.1.

Table 5.1: Taxi Fare-Rate Comparison Year 2014 -2015

Distance(km)	**Rate (Baht/km.) Year 2014**	**Distance(km)**	**Rate(Baht/km.) Year 2015**
First	35	First	35
2 - 12	Addition 5	1 - 10	Addition 5.5
12 - 20	Addition 5.5	10 - 20	Addition 6.5
20 - 40	Addition 6.0	20 - 40	Addition 7.5
40 - 60	Addition 6.5	40 - 60	Addition 8.0
60 - 80	Addition 7.5	60 - 80	Addition 9.0
More than 80	Addition 8.5	More than 80	Addition 10.5
Waiting Rate	1.5/minute	Waiting Rate	2.00/minute

The average fare-rate which calculate from our algorithm, the result shown that the new fare-rate increase from the old fare-rate about 13% as Figure 5.5.

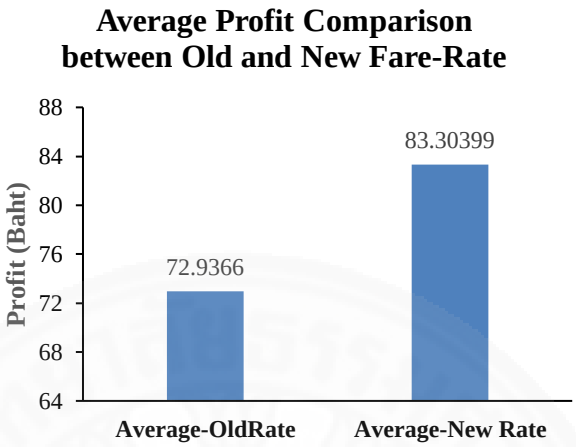

Figure 5.5: Average Total Cost Between the Old and New Fare-Rate

We also divide taxicab trip into mode as short distance journey, medium distance journey, and long distance journey as Table 5.2 from the graph in Figure 5.6.

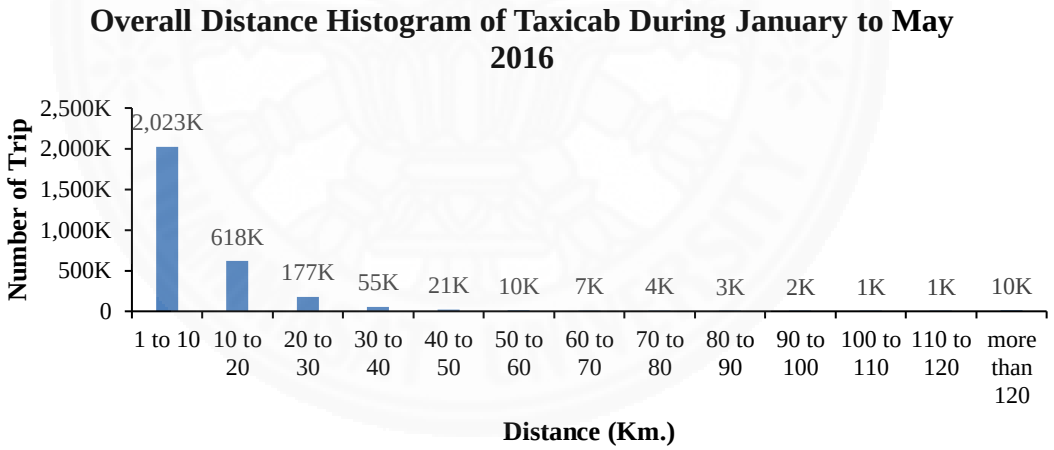

Figure 5.6: Taxi Distance Histogram

Table 5.2: Taxi Trip Distance Mode

Distance Mode	Distance in Kilometers
Long distance journey	More than 30
Medium distance journey	10 <= distance < 30
Short distance journey	1 <= distance < 10

When the new rate is applied, the fare-rate of each mode are shown in Figure 5.7. Where the increasing rate improve from the previous version of fare-rate as 14%, 17%, and 21% in short, medium, long distance journeys concurrently.

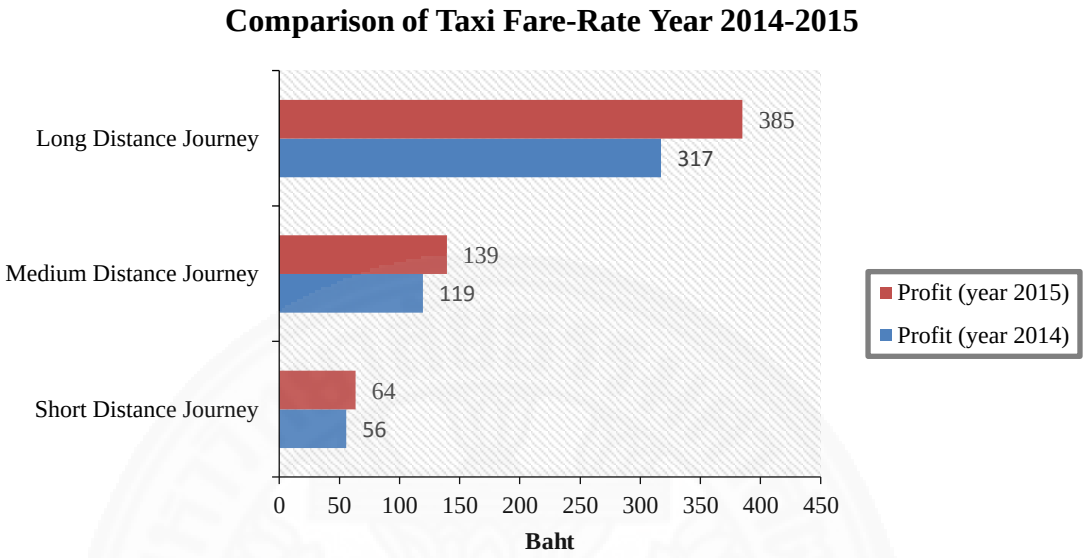

Figure 5.7: Comparison of Benefit when applied the New Fare-Rate on Distance Model

When we compared the usage of taxi which customer selected to their destination was considered as shown in Figure 5.8.

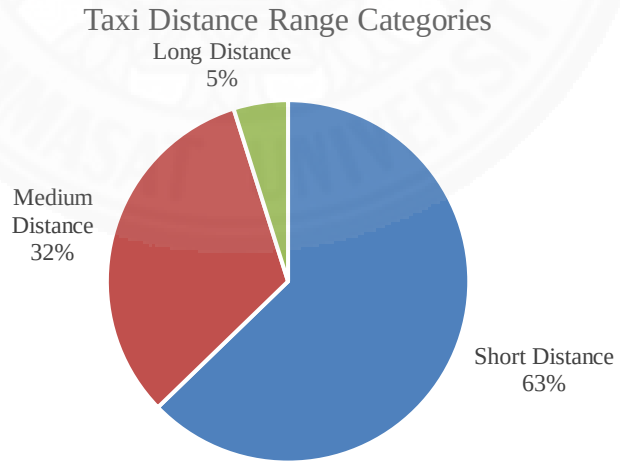

Figure 5.8: Trends of Taxi Trip in Bangkok

5.1.3 Which time are suitable for patrolling for customers?

The procedure for calculated the profit and valuable is as following diagram. We divide into 3 stages as Pre-Trip, Pick Up, and Trip Stage as Figure 5.9.

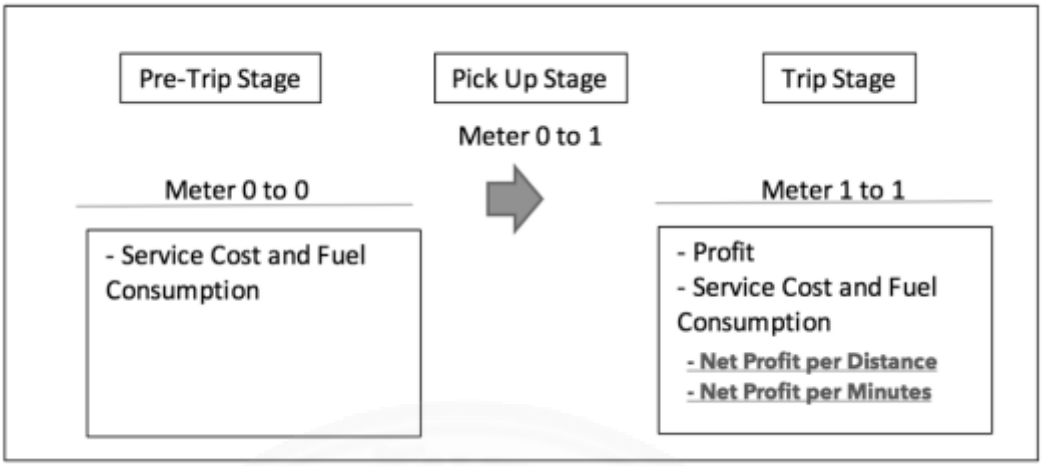

Figure 5.9: Taxi Trip Stage

In this study, we used Rama I and Thonburi as our case study to calculate taxi trip net profit and value. The area is shown as Figure 5.10 and Figure 5.11.

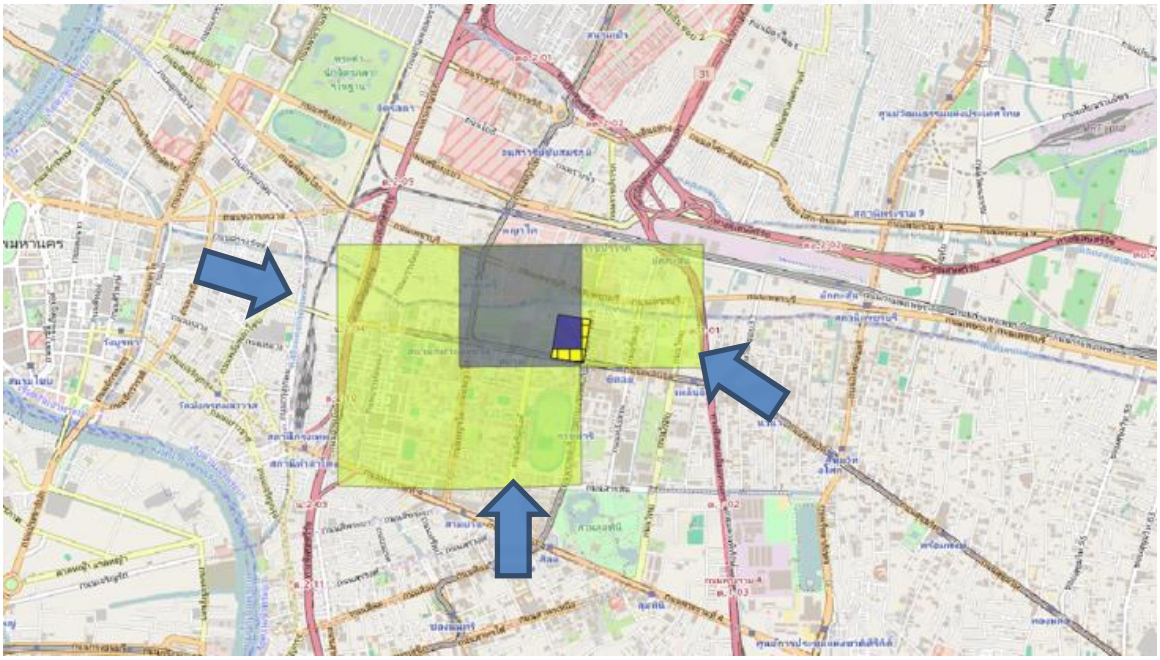

Figure 5.10: Rama I Bangkok

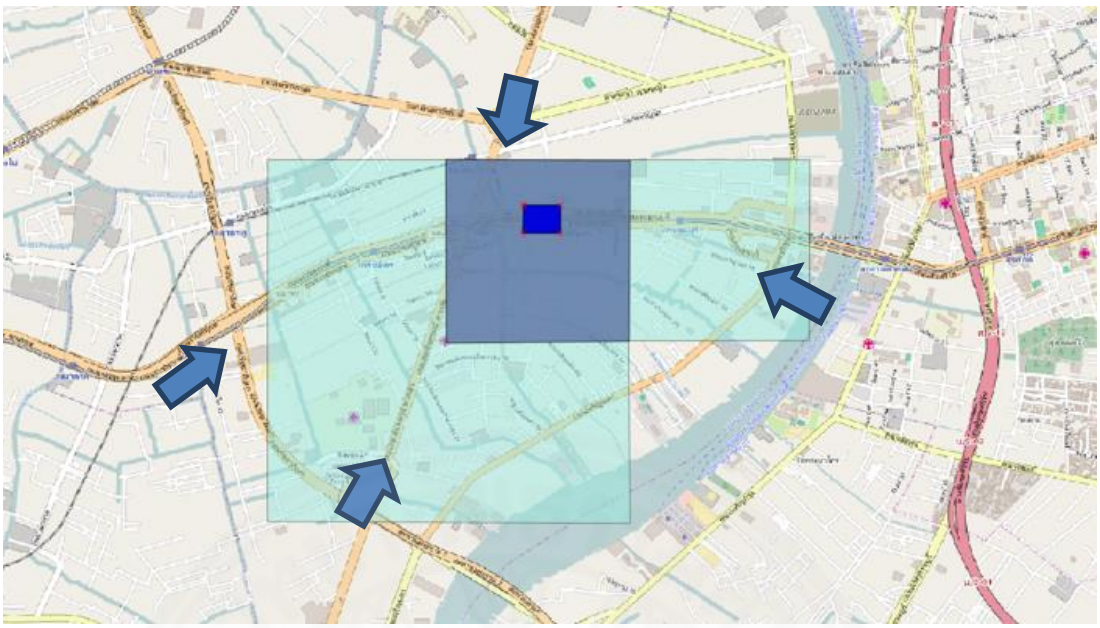

Figure 5.11: Thonburi, Bangkok

We discovered that in Figure 5.12, it is shown that profits would be increased directly relate to the operation work hour.

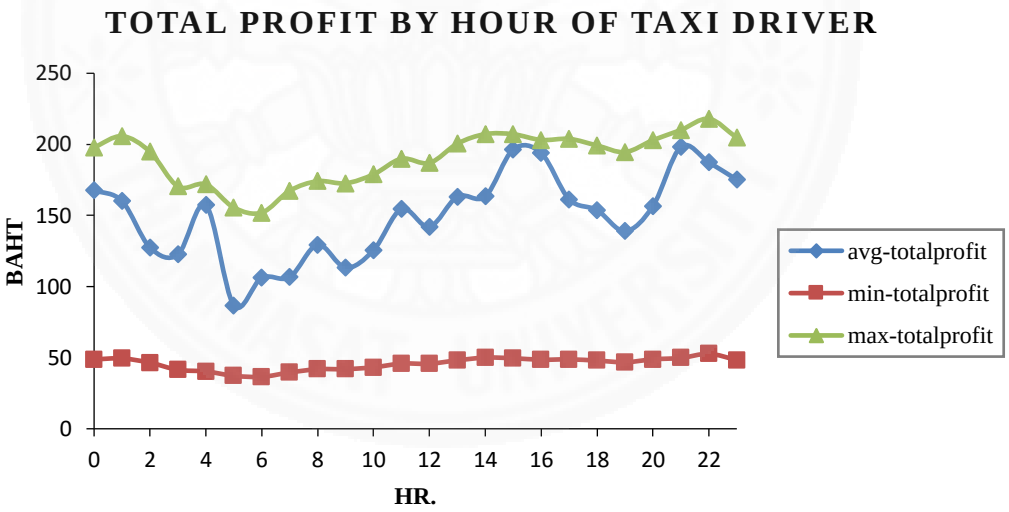

Figure 5.12: Total Profit when route from Rama I, Bangkok in Period of the day

Also, in Figure 5.13, is the total profit by hour when route from Thonburi, Bangkok.

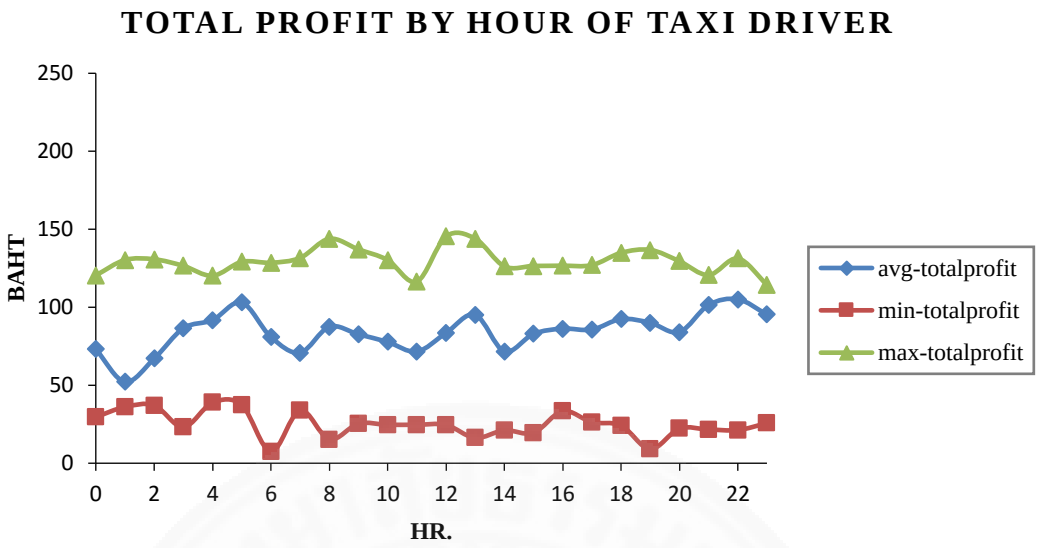

Figure 5.13: Total Profit when route from Thonburi, Bangkok

We also calculate the probability of taxi trip when route from Rama I and Thonburi. In Figure 5.14,we calculate probability of get customer when start trip from Central World Department Store, Rama I Bangkok.

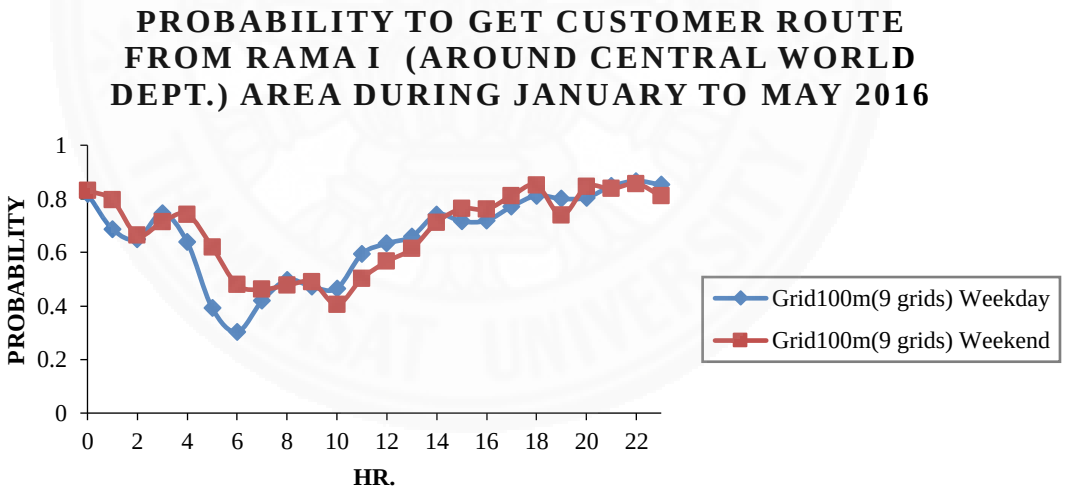

Figure 5.14: Probability to pick up customer when route from Rama I (Central World dept.)

We also apply probability calculation to Thonburi area as Figure 5.15.

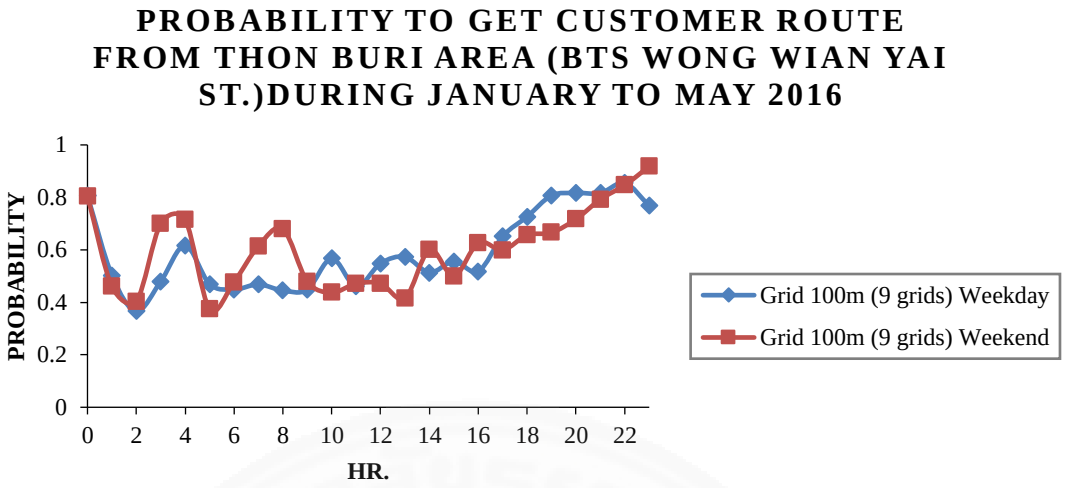

Figure 5.15: Probability to pick up customer from Thonburi (BTS Wong Wian Yai)

Finally, we have detected type of taxicab driver behavior when pick up customer. There are two main type which is detour (pick up along cruising) and parking (wait for customer in certain place). In Figure 5.16 is the data of taxicab from Rama I, Bangkok.

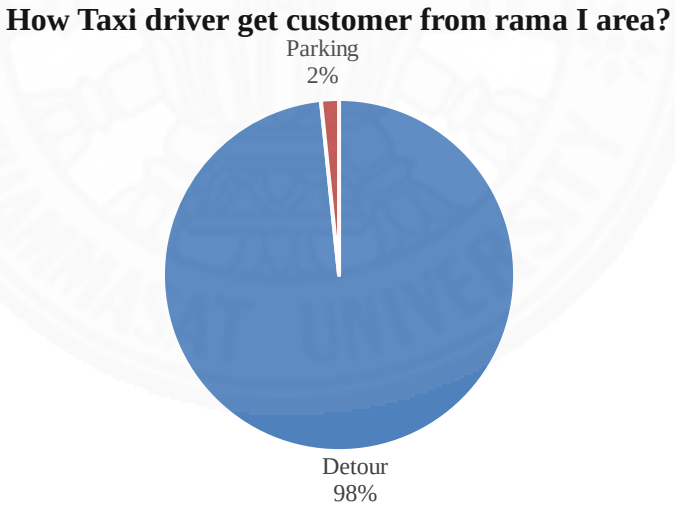

Figure 5.16: How taxi driver pick up customer? (From Rama I)

This will be varying depend on the area which we focus. In our survey, we get the result as shown in Figure 5.17.

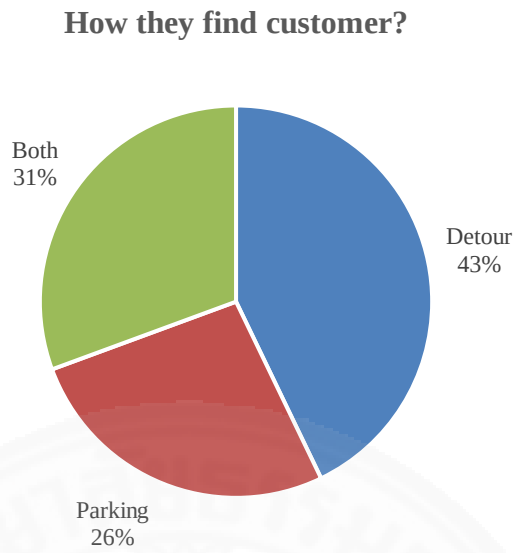

Figure 5.17: How taxi find customer? In our survey

5.1.4 Which type of routing can generate more profit?

We make comparison of distance mode. The result shown that run in short distance with more frequency is make taxi driver earn more profit than single long distance journey about 49.61% and compare with multiple medium distance make taxi driver earn more profit over long distance journey about 8.31% and Finally, we compare between short and medium distance and result show that run in multiple short distance will make increasing rate over medium distance about 38.13% shown in Figure 5.18.

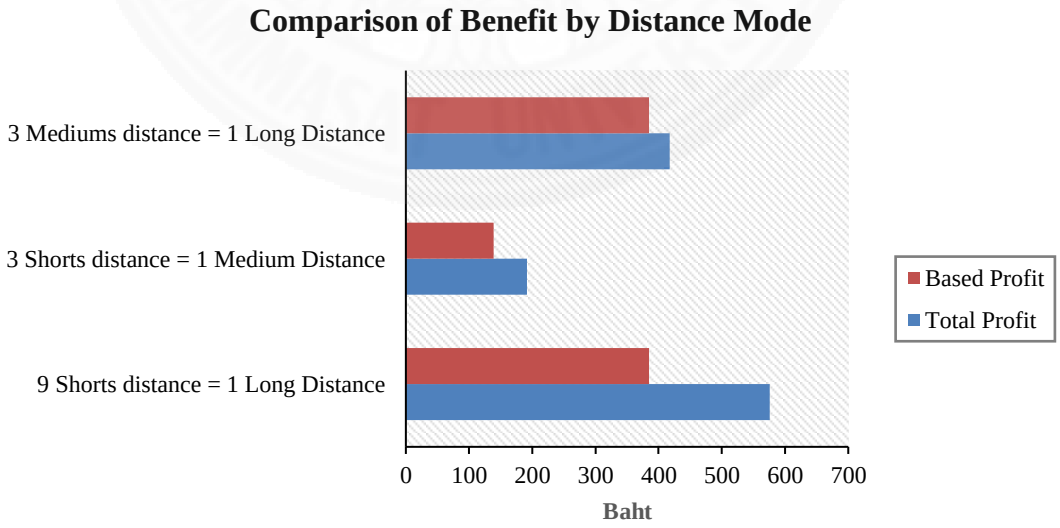

Figure 5.18: Comparision of Distance Mode

In another hand, we have make comparison on net profit per minutes of these travel modes. The result shown that some period short and medium distance give more net

profit per minutes than long distance mode such as 2.00 – 7.00 h. But in some period, Long distance is increased more in the afternoon as shown in Figure 5.19. The figure shown the result from Rama I, Bangkok. Another area will be different as shown in Figure 5.20.

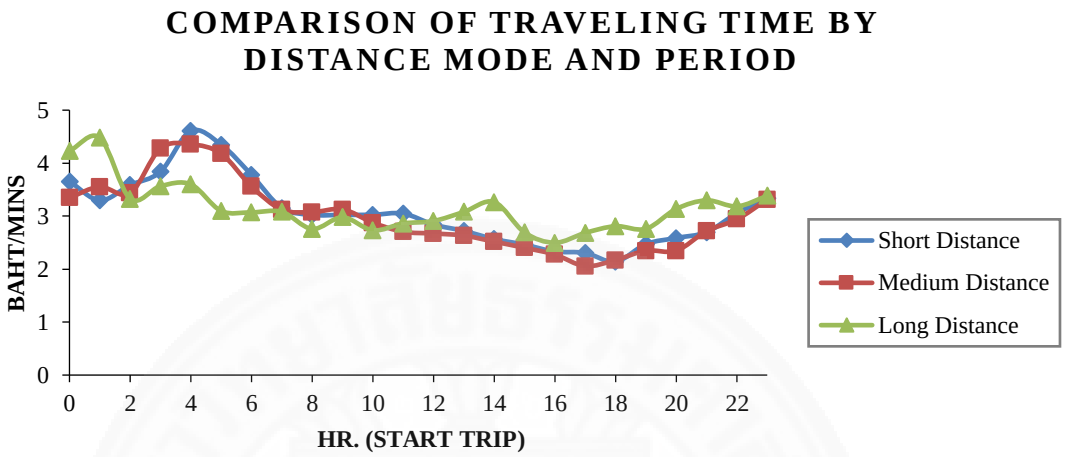

Figure 5.19: Net profit per Minute of distance mode

In Figure 5.20, shown how the area changed impact to the benefit from each distance mode some period run in short and medium obtained more benefit than long-distance in the late afternoon until midnight (15.00h. onward).

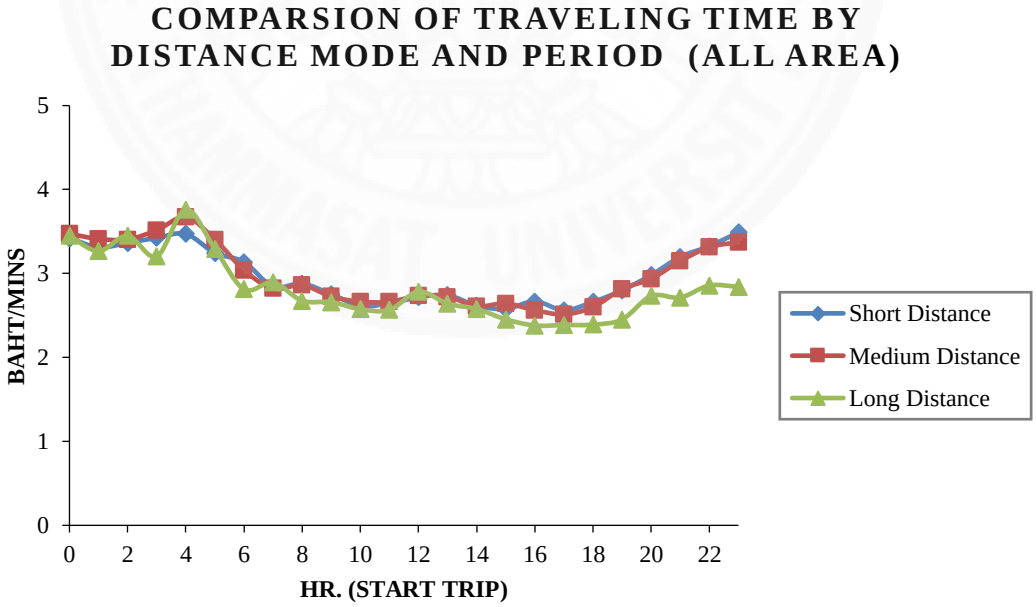

Figure 5.20: Net profit per Minute of distance mode (All Area)

5.1.5 How long does the taxi driver work per day?

In this section, we collect taxi driver's working hour from mobile application that we install on taxi driver's device. The duration of data collecting is about 15 days start from February 8 – 22, 2017. The results are as shown in Table 5.3.

Table 5.3: Taxi Working Hour Result

Content	Result
Departure Time	7.00 in average
Arrival Time	17.00 in average
Duration	14 hours per day in average
Population	9 taxi cabs
Duration of Collecting Data	31 Days

Then we make validation process with our paper-based survey that participant by 50 taxi drivers around Bangkok. From result in Figure 5.21 it shown that the working hour is most likely the same as from mobile application data since the most participant that install taxi data collecting application is a public taxi.

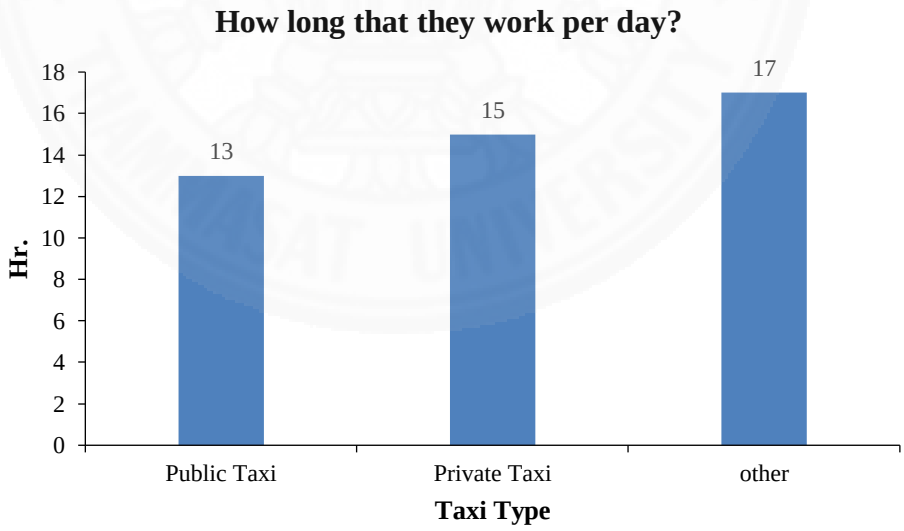

Figure 5.21: Taxi Working Hour from Paper-Based Survey

We also collect the cost that taxi driver's need to spent each time they enter the Gas Station as shown in Table 5.4.

Table 5.4: Cost at the gas station

Content	Result
Cost (Baht)	143.53 in average
Frequency per day	2.4 in average
Total Cost per day (Baht)	344.46 in average

Finally, we make visualize on taxi home, garage, and gas station as Figure 5.22.

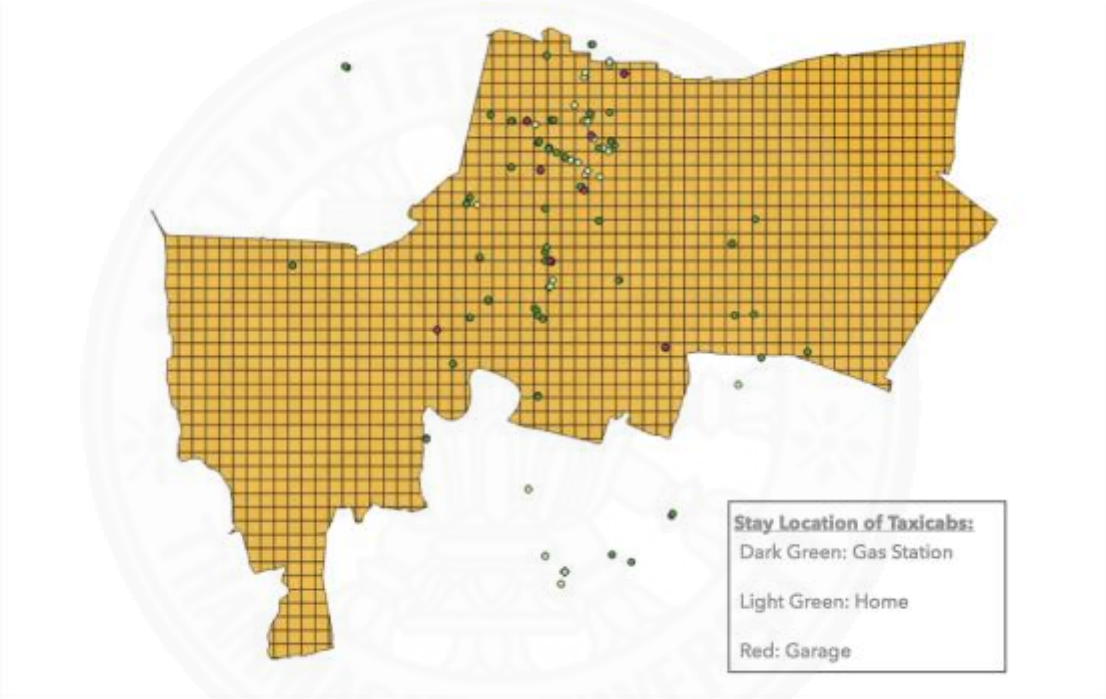

Figure 5.22: Taxi Stay Location

5.1.6 What is the future net-profit prediction model?

We apply Recursive Feature Elimination(RFE) method to select most importance feature as Figure 5.23.

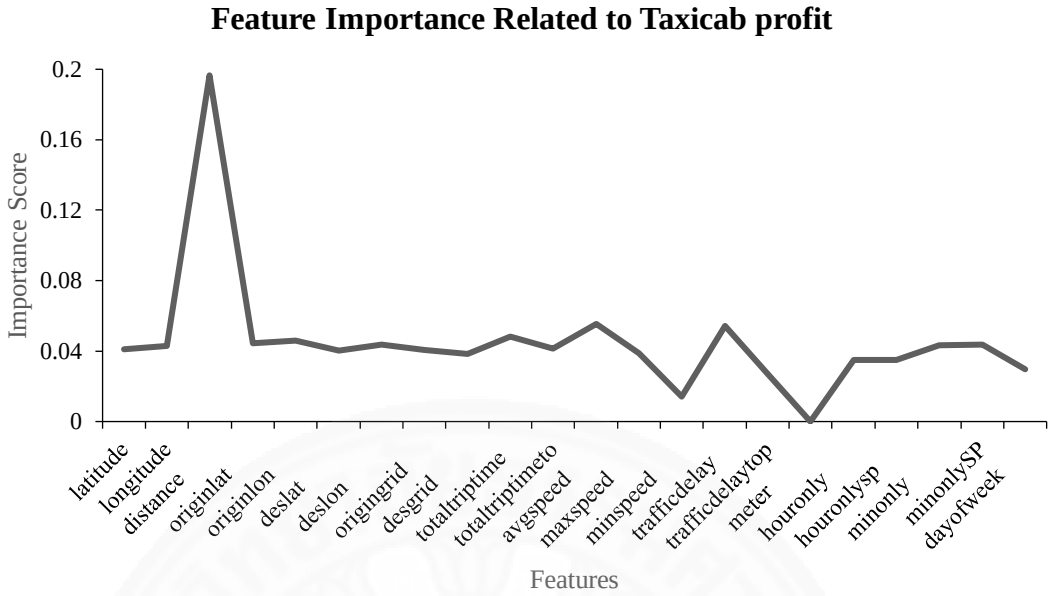

Figure 5.23: Feature Importance Related to Net-Profit

When we apply feature selection method we have built and obtained result from two models. One is predicted profit which build from our algorithm result and validate by taxi ground truth data (Real taxi trip) which we collected from android mobile application. The result is shown in Table 5.6 and Table 5.7.
The third model is the model that predict worthiness level of the taxi trip by classify trip net profit per distance into level by using jerk break method as in Table 5.5. The result is shown Table 5.10 and Table 5.11.

Table 5.5: Taxi Trip Worthiness Level

Range (Baht)	Level
More than 23.5	5
(12.9 – 23.5]	4
(7.8 – 12.9]	3
(4.2 – 7.8]	2
(-1 – 4.2]	1
Less than -1	0

Table 5.6: Profit Model Prediction Result

	Actual Profit	Predicted Profit	RMSE (Baht)	MAE (Baht)	Accuracy (%)	Error (%)	R-Squared
Model with Algorithm Result with 1,000 of training data							
Random Forest	123.27	135.15	11.88	5.37	90.36%	9.64%	0.96
Gradient Boost Tree	123.27	137.26	13.98	4.93	88.66%	11.34%	0.95
Decision Tree	123.27	135.57	12.30	4.73	90.02%	9.98%	0.96
Model with Algorithm Result with 5,000 of training data							
Random Forest	123.27	133.83	10.55	4.01	91.44%	8.56%	0.97
Gradient Boost Tree	123.27	135.46	12.19	3.74	90.11%	9.89%	0.96
Decision tree	123.27	135.59	12.32	3.34	90.01%	9.99%	0.96
Model with Algorithm Result with 10,000 of training data							
Random Forest	123.27	133.26	9.98	3.60	91.90%	8.10%	0.97
Gradient Boost Tree	123.27	135.02	11.74	3.30	90.47%	9.53%	0.96
Decision tree	123.27	134.32	11.04	3.15	91.04%	8.96%	0.97

Table 5.7: Profit Model Prediction Result (By Ground Truth)

	Actual Profit	**Predicted Profit**	**RMSE (Baht)**	**MAE (Baht)**	**Accuracy (%)**	**Error (%)**	**R-squared**
Random Forest	123.27	137.86	14.59	7.21	88%	12%	0.94
Gradient Boost Tree	123.27	141.10	17.83	8.14	86%	14%	0.92
Decision tree	123.27	141.13	17.85	8.14	86%	14%	0.92

In Table 5.8 and Table 5.9, we also validate the result by distance range by reduce the distance of testing data to 30 kilometers and 20 kilometers continuously. We use cross-validation to make the method select the best model for us to prevent overfitting from misconfiguration variable and bias.

Table 5.8: Model with Ground Truth (test with distance not more than 30)

	Actual Profit	**Predicted Profit**	**RMSE (Baht)**	**MAE (Baht)**	**Accuracy (%)**	**Error (%)**	**R-squared**
Random Forest	123.27	134.39	11.12	6.18	90.98%	9.02%	0.94
Gradient Boost Tree	123.27	137.48	14.21	6.96	88.48%	11.52%	0.90
Decision tree	123.27	137.48	14.21	6.96	88.48%	11.52%	0.90

Table 5.9: Model with Ground Truth (test with distance not more than 20)

	Actual Profit	Predicted Profit	RMSE (Baht)	MAE (Baht)	Accuracy (%)	Error (%)	R-squared
Random Forest	123.27	133.07	9.80	5.51	92.05%	7.95%	0.91
Gradient Boost Tree	123.27	135.23	11.96	6.01	90.30%	9.70%	0.87
Decision tree	123.27	135.23	11.96	6.01	90.30%	9.70%	0.87

Also, we have compare model compute performance of Random forest, Gradient Boost Regression Tree and Decision Tree as shown in Figure 5.24.

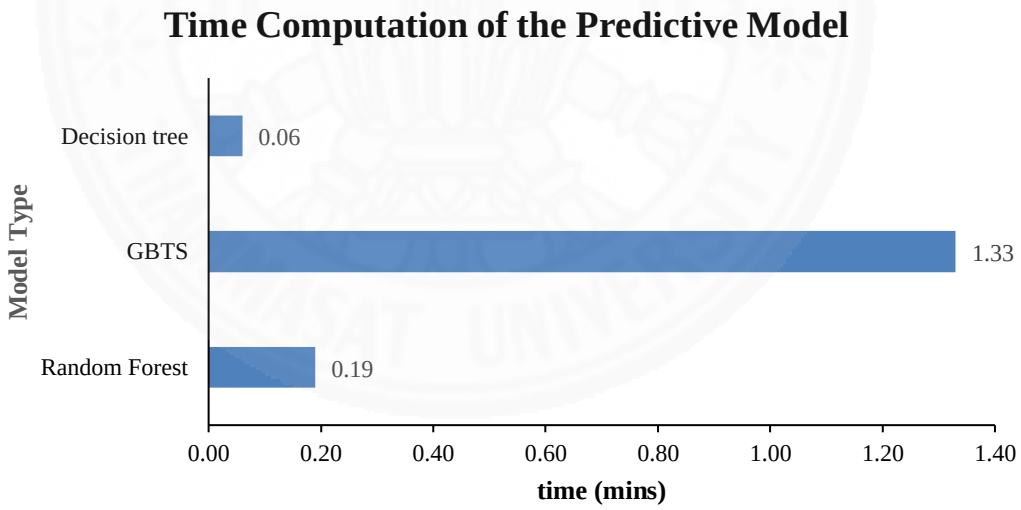

Figure 5.24: Comparison on time computation of predictive model

Table 5.10: Worthiness Classification (Training Stage)

Number of Tree	Error	Accuracy
5	10.32%	89.68%
10	9.15%	90.85%
50	8.37%	91.63%

Table 5.11: Worthiness Level Classification (Test against all data)

	Error	Accuracy
Test Against All Data	10.57%	89.43%

We also evaluate the result by using confusion matrix as shown in Table 5.12.

Table 5.12: Confusion Matrix on Multi classification method

Class	**0**	**1**	**2**	**3**	**4**	**5**	**Average**
TP	0.87	0.85	0.91	0.86	0.83	0.73	0.84
FP	0.01	0.05	0.09	0.04	0.02	0.00	0.03
TN	0.13	0.15	0.09	0.14	0.17	0.27	0.16
FN	0.99	0.95	0.91	0.96	0.98	1.00	0.97

5.1.7 How much fee do we need to increase to earn drivers more profits?

In this section, we created a mathematic model, to calculate amount of fee to add on top of regular fare-rate to make drivers earn more profit. In Table 5.13, we analyze from our data and list profit loss per distance into percentage, profit loss per minute, and probability when this profit loss happens.

Table 5.13: Profit loss and Probability

Loss	**Lost (Baht per km.)**	**Lost (Baht per min.)**	**Probability**
Less than or equal to 10%	3.50	1.31	0.01
10% to 20%	3.33	1.62	0.08
20% to 30%	3.39	1.76	0.25
30% to 40%	4.38	1.97	0.23
40% to 50%	5.79	2.44	0.14
50% to 60%	7.27	3.03	0.08
60% to 70%	8.81	3.65	0.05
70% to 80%	10.43	4.28	0.04
80% to 90%	12.06	4.94	0.03
90% to 100%	13.84	5.54	0.02
more than 100%	26.13	9.83	0.06

We also comparison between solution which taxi driver propose and our developed solution as Table 5.14 and Table 5.15.

Table 5.14: Propose solution from taxi driver's perspective

No.	Solution
1	Route to airport add 50 baht
2	Peak hour adds 50 baht
3	Reservation by application add 50 baht

Table 5.15: Propose solution from our perspective

No.	Solution
1	Probability Less than 0.2 in destination
2	Peak Hour add more by distance range
3	Average Speed Less than 20 km/h add more

In our solution, we focused on probability of the destination, period, and the speed of the taxi vehicle and add on top fare by distance as Table 5.15. But in the taxi driver's proposed solution, they used fix on top fare so it causes advantage to all the drivers but make disadvantage to the customer. We will make discussion on this in the discussion section. In Table 5.16, is the example of fare adjustment which we used to calculate our solutions. Also, in Table 5.17 which we used calculation result to make suggestion in distance range.

Table 5.16: Example of Fare Adjustment Calculation

	real(non-traffic) trip	**real(traffic) trip**	**average trip**	**ideal trip**
Actual Profit	57	57	57	57
Net Profit	46	34	40	51
Solve Profit	64	79	72	57
Differentiate	0.20	0.40	0.30	0.10
Additional Charge	7	22	15	0

Table 5.17: Example on Fare Adjustment Calculation by Distance Range

Distance	**Add(Baht)**
0 to 10	15
10 to 20	30
20 to 30	45

5.2 Result Discussion

In the beginning of our research, we have started with data exploration on the raw data and make unwanted data cleaning then we constructed an algorithm to handle taxi probe data to calculate it taxicab's trip basic statistic such as distance, speed, trip-time, traffic delay, and actual trip cost. Our research result has been divided into 6 sections. So, in this section we are make discussion of them.

First, in section 5.1.1 Paper-Based Taxi Survey, We have begun this research result with a taxi paper-based survey to get drivers general information. From this process got about 58 taxi drivers to be our sample, we have 58% is private taxi, who own a car to make a service, 39% is public taxi where the driver rent the car from a garage and 3% is other type of taxi where not match these categories. We have discovered that famous place to pick up customer is the location where economic and department store are located, the second rank is transportation area, and third is the location where school and university are located. Since the main of this study is to find out the

answer of the reason “why the taxi decline service on customer in Bangkok”. The survey shown that 62% of our sample have experience in decline service to customer and the most reason that they decline service is that customers ask them to go to the place which opposite or not in the desire route of the taxi drivers. The customer will not satisfy with is result.

Next, in section 5.1.2 How does new taxi fare-rate gain taxi driver to gain more *benefit*, When the new rate is applied, we discovered that the new fare rate will make benefit to the taxi driver if they are in the long-distance route and the increasing rate was about 21% (68 Baht) and medium-distance route brought about an increasing rate by 17% (20 Baht) and 14% (8 Baht) in benefit when the short-distance route is used. Then When we compared the usage of taxi which customer selected to their destination was considered as shown in Figure 5.8. It is demonstrated that their might explain why they choose to use taxi service in the short-distance route more than the long-distance route so therefore most of the taxi drivers need to increase rate per kilometer to be higher than the present rate. From this point, we could not summarize yet if the fare-rate need to be improve for the short distance range as taxi-driver requested. This need to be support by another factor which will be describe in a further section.

Then in section 5.1.3 Which time are suitable for patrolling for customers?, We have selected area of Rama I, and Thonburi, Bangkok as our study area. From the result, has shown that profit increase during 5.00 h until 15.00 h and steady on 16.00 h then start to drop on the peak hour period (17.00 – 19.00 h.). We would recommend taxi driver that on 4.00, 15.00, 16.00, and 21.00 h. are the suitable hour which has a chance to get high profit. But for in Thonburi area we would recommend taxi drivers to starting routing in the early morning (1.00 – 5.00 h.). Also, the probability to get customer from Rama I will increasing due to the operating hour of the department store. But for Thonburi will we peak after 16.00 h. until midnight. In Rama I, Taxi drivers get customer by detour rather than parking and wait for them. This also can be change due to the area and the environment at that location. Next is section 5.1.4 Which type of routing can generate more profit?, we could demonstrate that the distance of taxicab trip is so importance to the benefit that taxi driver will get in each trip. We make comparison between short trip and long trip and show that frequent short trip give more opportunities to earn more profit than run in single long distance with the same amount of distance. To make the performance of taxi driver increase and earn more benefit, they should concern about the destination they willing to go. In the past, they think that run in the long distance will make them earn more profit but instead run in the shorter distance more frequent is make sense to make them earn more profit and save their expense-cost. The traffic and distance of pre-trip are needed to concern too. Also, minimize the service cost for find customer is increasing chance to have higher performance of taxi driver to earn more profit and the expense-cost are reduced too. In this section, we have compare between net-profit and time spent and shown that for Rama I area between 2.00 – 6.00 h. is the suitable time for taxi to drive in short or medium distance than in long distance mode. Then in the middle of the day there are nearly the same between distance modes. But in the afternoon long distance provide more profit to drivers. This may cause by the traffic condition in the central of the city. This result may change due to the area of interest.

After that in section 5.1.5 How long does the taxi driver work per day?, we used mobile application to collect data such as meter information for our model validation. We also collect home, garage, and gas location. Which we discover that taxi driver or our survey usually work at least 14 hours per day and make gas refill at the gas station twice a day. The cost of gas the need to pay is about 344.6 baht in average.

In section 5.1.6 What is the future net-profit prediction model, the result of modelling testing show that Random Forest regression is suitable for modeling tuning when compare to another model from B.1 Root Mean Square Error(RMSE) Comparison. after tuning and data preprocessing on 1.5 million taxicabs trips in Bangkok and valid with our data ground truth. The model perform prediction only 10 Baht error. The prediction gives opportunities to predict and recommendation for the future travelling trip. Also, shown that our data is accurate over 90 percent. The model also predicts the accurate result when we input tested data in the shorter trips because of our training data did not contain taxicab trip that run in the long distance (more than 50 kilometers) much so it could degrade in prediction result. So, the trip in the short run will have higher prediction accuracy result than in the long run trip. We have discovered that Gradient Boost Regression Tree use time to perform model training more that another two model (Random forest and Decision Tree) in our model testing. Also, when we increase the data size of training data it increases the computational time as shown in Figure B.2. We also have model that predict worthiness level of the taxicab's trip by input features into it. The prediction give a high-performance result which got over 90 percent accuracy. From our data in Figure B.1 show that most of the taxi trip is in level 2 where net profit per distance is between 4.2 -7.8. The outcome could be based model to use for taxi applications when they input and want to know how worthiness of routing and which driving behavior is recommend in the real situation.

Finally, we have come up with solutions that could make taxi's driver earn more profit. We have comparison on both solution that has been proposed by taxi drivers and our proposed solution. We have discovered that the proposed solution from taxi side make benefit directly to the drivers but it not satisfies to the customer. For example, if we run in short distance say 5 kilometers and have traffic delay 20 minutes. In the old way, we paid 77 baht but if we follow the new solution of the taxi drivers as Table 5.14 we paid 127 baht, which has 65% increased from the old one. This make customer not satisfy with it. In another hand, if used our solution as Table 5.15 with the same criteria. Customers will pay only 92 Baht which increased from the old one. This solution make satisfy on both customer and drivers. The driver could make more profit and customer do not to have to pay high fare.

Chapter 6

Conclusions and Recommendations

6.1 Research Summary

To conclude our research, we have start with data exploration to discover data pattern and basic statistic then we developed an algorithm to handle taxi probe data and calculate taxi total fare-rate of each trips. We also built a model for calculate taxi net profit by take input from our algorithm result. Our first finding, we discover that Taxi driver in our sample about 62% has experience on decline service to customer and the has the most reason is that the route which has been request is different than their desire route. Then our second finding is that new fare-rate which apply in year 2015 make benefit about 13 percent increasing from the old version and make most impact to the long-distance journey trip. The government try to support taxi where they run in long journey. Instead we also found that most of taxicabs run short distance journey more often than the long journey. The short journey also makes more profit than the long-distance journey when use the same amount of distance range (multiple short trip compare with single long trip) Second, the probability of get customer also depends on the area of interest. The area which are well-known such as department store or city center have high probability to occur taxi trip. Then we also discover that most of taxi in Rama I, Bangkok Thailand love to detour rather than parking in some place to get customer. This also depend on the area. Third, Taxi working hour is at least 14 hours per day. Finally, in taxi prediction model, we discovered that Random Forest Regression Tree is the easiest to make model tuning and give net-profit prediction which has only 10 Baht error. And the result accuracy is over to 80 percent in trip's profit prediction and worthiness level. We also come up with on top fare-rate suggestion on amount of fare-rate should be increase to make taxi drivers earn more profit and make customer to get more ride to the destination

To make suggestions to taxi drivers in Bangkok, from our research finding, we would suggest that taxi driver need to concern on these factors. First, location where they run to get customer (Shorter run is saved expense cost). Second, distance of routing trip. The more shorter distance make more profit than the long distance when use the same range of distance. Finally, the time of routing (driving during peak hour is impact directly to taxi profit which cause mainly from traffic situation). We come up with solution to make taxi driver earn more profit and reduce decline service to customer by adding fare-rate to regular fare to make drivers earn more profit. In distance, less than 10 kilometers will add fare-rate 15 baht, 10 to 20 kilometers will add 30 baht and more than 20 kilometers will add 45 baht.

In our future work, we will expand study area to be cover most of the area in Bangkok since in this research we used Rama I and Thonburi, Bangkok as our case study.

6.2 Key Contribution of the Research

6.2.1 Data Exploration

This method use for study pattern and basic statistic of our data. In this research, we used geographic information knowledge such as intersection and buffering geospatial polygon to cluster data into area and clean all outliner which not located in the road network. In the future work, we recommend using the road segment as the data clustering method. It could use to detect and cluster GPS data point more efficient in term of speed detection since we divide road into segments.

6.2.2 Taxi Trip Assessment Model

This framework use to calculated and evaluated taxi trip worthiness. It combines between mathematic model and large scale geospatial data to make analysis on trajectory cost and value. This model could use to suggest taxi trip profit and worthiness value and make useful information for taxi driver to plan for their driving habit. Also, how much on top fare-rate should be add to make drivers earn more profits. For example: Should run shorter to save cost and increase more profit? Or Continue to run in long distance? We would recommend this model to use with embedded system in taxi vehicle and act as smart taxi in the future.

6.2.3 Data Modeling and Evaluation

The model which we have built contain of two model. First, is the model which predict and evaluate taxi's profit that are result from our mathematic model. The second, the model which predict taxi's trip worthiness value. In both model, we input features such as distance, time, delay, hour when trip start that related to taxi profit. We build the model and make prediction. The evaluate result of our result are over 80 percent accuracy which valid by data ground truth. We would recommend that to increasing evaluation accuracy can be done by have a method that increasing accuracy of speed detection in the ground truth which we used mobile application to collected data.

References

Ashbrook, D. and Starner, T. (2003) Using GPS to learn significant locations and predict movement across multiple users, *Personal and Ubiquitous Computing*, 7 (5), pp. 275–286. DOI:10.1007/s00779-003-0240-0.

Bai, Y. and Wang, E. (2012) Design of Taxi Routing and Fare Estimation Program with Re-prediction Methods for a Smart Phone.

Bondy, J. A. and Murty, U. S. R. (2008) *Graph Theory*. Vol. 244. London: Springer London. DOI:10.1007/978-1-84628-970-5.

BYON, Y. J. and LIANG, S. (2014) Real-Time Transportation Mode Detection Using Smartphones and Artificial Neural Networks: Performance Comparisons Between Smartphones and Conventional Global Positioning System Sensors, *Journal of Intelligent Transportation Systems*, pp. 264–272.

Ding, Y., Liu, S., Pu, J. and Ni, L. M. (2013) HUNTS: A trajectory recommendation system for effective and efficient hunting of taxi passengers, *Proceedings - IEEE International Conference on Mobile Data Management*, 1, pp. 107–116. DOI:10.1109/MDM.2013.21.

Doa, T. M. T. and D. G.-P. (2013) Where and what: Using smartphones to predict next locations and applications in daily life., in: *Pervasive and Mobile Computing*. pp. 79–91.

Egan, M. and Jakob, M. (2016) Market mechanism design for profitable on-demand transport services, 89, pp. 178–195. DOI:10.1016/j.trb.2016.04.020.

Etter, V. and M. K. (2013) Where to go from here? Mobility prediction from instantaneous information., *Pervasive and Mobile Computing*, pp. 784–797.

Granitto, P. M., Furlanello, C., Biasioli, F. and Gasperi, F. (2006) Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products, *Chemometrics and Intelligent Laboratory Systems*, 83 (2), pp. 83–90. DOI:10.1016/j.chemolab.2006.01.007.

He, Z., Yang, L. and Wei Guan (2014) A day-to-day route choice model based on travellers' behavioural characteristic, in: *International Conference on Traffic & Transportation Studies (ICTTS'2014)*.

Hwang, R.-H., Hsueh, Y.-L. and Chen, Y.-T. (2015) An effective taxi recommender

system based on a spatio-temporal factor analysis model, *Information Sciences*, 314 (4), pp. 28–40. DOI:10.1016/j.ins.2015.03.068.

James, G., Witten, D. and Tibshirani, T. H. R. (2015) *An Introduction to Statistical Learning with Applications in R, Performance Evaluation*. Vol. 64. DOI:10.1007/978-1-4614-7138-7.

Kamimura, J., Ogawa, M., Wakayama, H., Iga, N., Shiota, N. and Yano, M. (2013) D-Taxi: Adaptive area recommendation system for taxis by using DiRAC, *2013 International Conference on Connected Vehicles and Expo, ICCVE 2013 - Proceedings*, (4), pp. 507–508. DOI:10.1109/ICCVE.2013.6799845.

Liaw, a and Wiener, M. (2002) Classification and Regression by randomForest, *R News*, 2 (December), pp. 18–22. DOI:10.1177/154405910408300516.

Lin, M. and H., W.-J. (2014) Mining GPS data for mobility patterns: A survey, *Pervasive and Mobile Computing*, pp. 1–16.

Manley, E. J., S.W. Orr and Cheng., T. (2015) A heuristic model of bounded route choice in urban areas., *Transportation Research Part C*, pp. 195–209.

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J. and Damas, L. (2013) Predicting Taxi #x2013;Passenger Demand Using Streaming Data, *Intelligent Transportation Systems, IEEE Transactions on*, 14 (3), pp. 1393–1402. DOI:10.1109/TITS.2013.2262376.

Nilsson, N. J. (2005) Introduction to Machine Learning, *Machine Learning*, 56 (2), pp. 387–99. DOI:10.1016/j.neuroimage.2010.11.004.

Noulas, A., Scellato, S., Lathia, N. and Mascolo, C. (2012) Mining User Mobility Features for Next PlacePrediction in Location-based Services, in: *ICDM'2012*.

Phiboonbanakit, T. and Horanont, T. (2016) Who will get benefit from the new taxi fare rate? Discerning the real driving from Taxi GPS data, in: *2016 7th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*. IEEE, pp. 73–78.

Qi, G., Pan, G., Li, S., Wu, Z., Zhang, D., Sun, L. and Yang, L. T. (2013) How long a passenger waits for a vacant taxi? Large-scale taxi trace mining for smart cities, *Proceedings - 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, GreenCom-iThings-CPSCom 2013*, (4), pp. 1029–1036.

DOI:10.1109/GreenCom-iThings-CPSCom.2013.175.

Qu, M., Zhu, H., Liu, J., Liu, G. and Xiong, H. (2014) A cost-effective recommender system for taxi drivers, *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '14*, pp. 45–54. DOI:10.1145/2623330.2623668.

Salanova, J. M., Estrada, M., Aifadopoulou, G. and Mitsakis, E. (2011) A review of the modeling of taxi services, *Procedia - Social and Behavioral Sciences*, 20, pp. 150–161. DOI:10.1016/j.sbspro.2011.08.020.

Scellato, S., Musolesi, M., Mascolo, C., Latora, V. and Campbell, A. T. (2011) NextPlace: A Spatio-temporal Prediction Framework for Pervasive System, in: *Pervasive 2011*. pp. 152–169.

Somkiadcharoen, D., Horanont, T., Pattara-atikom, W. and Sugino, N. (2015) Data Exploration of Taxi during Protesting Period in Thailand by GPS Tracking, in: *Proceedings of the Tenth International Conference on Knowledge, Information and Creativity Support Systems (KICSS 2015)*. pp. 493–504.

Terry, F. (2014) Taxi fares go up in Bangkok Saturday. Available from: http://www.bangkokpost.com/learning/learning-news/449490/taxi-fares-go-up-in-bangkok-today [Accessed

Thomas, T. and Tutert, B. (2015) Route choice behavior in a radial structured urban network: Do people choose the orbital or the route through the city center?, *Journal of Transport Geography*, pp. 85–95.

Witayankurn, A., Horanont, T., Ono, N., Sekimoto, Y. and Shibasaki, R. (2013) Trip Reconstruction and Transportation Mode. International Conference on Computers in Urban Planning and Urban Management,

Yingjun, Y., Cui, H., Shaoyang, Z. and Yingjun, Y. (2012) A prediction Model of the Number of Taxicabs Based on Wavelet Neural Network *, 12 (Icese 2011), pp. 1010–1016. DOI:10.1016/j.proenv.2012.01.380.

Yuan, J., Zheng, Y., Zhang, L., Xie, Xi. and Sun, G. (2011) Where to Find My Next Passenger, pp. 109–118. DOI:10.1145/2030112.2030128.

Yue, Y., Zhuang, Y., Li, Q. and Mao, Q. (2009) Movement Patterns from Taxi Trajectory Data, *Knowledge Creation Diffusion Utilization*.

Zhang, D. and He, T. (2012) PCruise: Reducing cruising miles for taxicab networks,

Proceedings - Real-Time Systems Symposium, pp. 85–94. DOI:10.1109/RTSS.2012.61.

Zhang, M., Liu, J., Liu, Y., Hu, Z. and Yi, L. (2012) Recommending pick-up points for taxi-drivers based on spatio-temporal clustering, *Proceedings - 2nd International Conference on Cloud and Green Computing and 2nd International Conference on Social Computing and Its Applications, CGC/SCA 2012*, pp. 67–72. DOI:10.1109/CGC.2012.34.

Zhang, Y. and Haghani, A. (2015) A gradient boosting method to improve travel time prediction, *Transportation Research Part C*, 58, pp. 308–324. DOI:10.1016/j.trc.2015.02.019.

Zheng, Y., Zhang, L., Xie, X. and Ma, W.-Y. (2009) Mining interesting locations and travel sequences from GPS trajectories, in: *Proceedings of the 18th international conference on World wide web - WWW '09*. pp. 791.

Zhou, C., Dai, P., Wang, F. and Zhang, Z. (2016) Predicting the passenger demand on bus services for mobile users, *Pervasive and Mobile Computing*, 25 (2013), pp. 48–66. DOI:10.1016/j.pmcj.2015.10.003.

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S. and Terveen, L. (2007) Discovering personally meaningful places, *ACM Transactions on Information Systems*, 25 (3), pp. 12–es. DOI:10.1145/1247715.1247718.

Appendices

Appendix A

Function and Algorithm

The overall function in taxi fare-rate calculation algorithm Chapter 3: Methodology are described as the following description

A.1 Taxi Fare Rate Calculation Algorithm

```
Initialize cost, totaltriptime , traffic delay and distance to zero
Initialize Array for speed
Initialize p_meter, p_grid, p_lat,p_lon and p_dt to "None"
While data I less than total data size
    If p_meter is equal to "None" or "0" and meter is equal to meter is equal to "1"
        Set p_meter to meter,Set p_grid to grid, Set p_dt to dt
        Set p_lat to lat and olat to lat, Set p_lon to lon and olon to lon
    Else if p_meter is equal to "1" and meter is equal to meter is equal to "1"
    Set distance equal the sum of current distance and distance from distance calculation function
    Set Total trip time equal the sum of current totaltriptime and time from findtime function
    If speed less than 6
        Set traffic delay equal the sum of current trafficdelay and time from findtime function
    Input speed to array Speed
    Set p_meter to meter,Set p_grid to grid,Set p_dt to dt,Set p_lat to lat
    Set p_lon to lon
    Else if p_meter is equal to "1" and meter is equal to meter is equal to "0"
    Set distance equal the sum of current distance and result from distance calculation function
    Set Total trip time equal the sum of current totaltriptime and result from findtime function
    Set cost equal the input cost and traffic delay to findcost function
    Set dlat to lat and dlon to lon
    Print imei,lat,lon,olat,olon,dlat,dlon,distance,totaltriptime,traffic delay,cost,dt
```

This algorithm use to calculate taxi trajectory movement such as distance, average speed, trip time, traffic delay and profit which obtained from each trip.

A.2 Stay place Detection Algorithm

```
Initialize p_speed, p_grid, p_lat,p_lon and p_dt to "None"
While data I less than total data size and meter is equal to zero
If p_speed is more than zero and speed is equal to zero
        Set p_speed to speed Set p_grid to grid Set p_dt to dt Set p_lat to lat and olat to lat
        Set p_lon to lon and olon to lon Set dtin as dt
Else if p_speed is equal to zero and speed is equal to "0"
        Set time equal the sum of current time and time from findtime function
        Set p_speed to speed Set p_grid to grid Set p_dt to dt Set p_lat to lat Set p_lon to lon
Else if p_speed is equal to zero and speed is more than zero
        Set time equal the sum of current totaltriptime and result from findtime function
        Set p_speed to speed Set p_grid to grid Set p_dt to dt Set p_lat to lat  Set p_lon to lon
        Set dlat to lat , dlon to lon and dtout to dt
        Print imei,lat,lon,olat,olon,dlat,dlon,p_grid,grid,time,dtin,dtout
```

The stay place detection algorithm use to detect home/gas/garage location. It also used to define how taxi driver pick up customer (Detour or Parking)

A.3 Taxi Stop Interval State

The state is analyzed, if less than 6 Hrs. we detect it as the short stop and within 6 Hrs. and stop at least 10 minutes we detect it as park state otherwise it is detour state (Pick up customer along the way). If stop more than 6 Hrs. we detect it as home or garage or gas station depend on frequency of visited

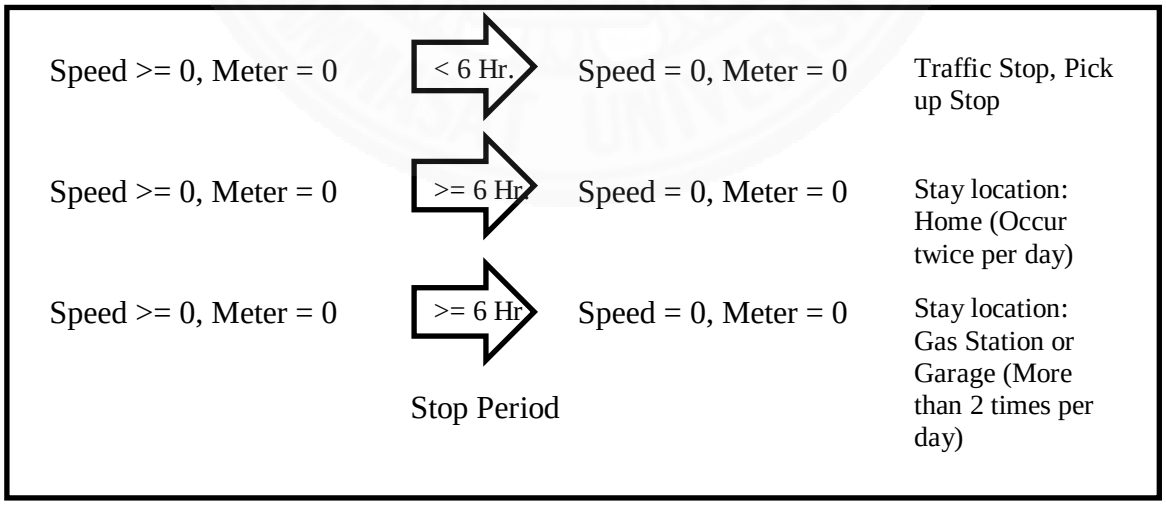

Figure A.1 Taxi Stop State

A.4 Distance Calculation Function

This function used to calculate distance between two point of taxi trajectory movement.

```
def distanceRoute(x,y):
    set c_lat equal to x
    set c_lon equal to y
    set a1 equal to p_lat multiply by rad
    set a2 equal to p_lon multiply by rad
    set b1 equal to c_lat multiply by rad
    set b2 equal to c_lon multiply by rad
    set dlon equal to b2 subtract a2
    set dlat equal b1 subtract by a1
    a = (math.sin(float(dlat)/2))**2 + math.cos(a1) * math.cos(b1) *
    (math.sin(float(dlon)/2))**2
    c = 2 * math.atan2(math.sqrt(a), math.sqrt(1 - a))
    R = 6378.145
    distance = (float(R) * c)
    return distance
```

A.5 Time Calculation Function

This function use to calculate time different between two GPS data points and sum up as the total trip time and traffic delay

```
def findtime(minute):
    if previous min less than minute:
        Set time equal to 60 subtract by previous minute then add with current minute
        return time
    else:
        time equal to minute subtract by previous minute
    return time
```

A.6 Findcost calculation Function

This function used to calculate taxi trip cost between the old and new fare-rate by use input from distanceRoute and findtime function

```
Initialize CostNew to zero
Initialize distancepercurrency1, distancepercurrency2, distancepercurrency3,
distancepercurrency4, distancepercurrency5 and distancepercurrency6 to zero
Initialize distance1, distance2, distance3, distance4, distance5 and distance6 to zero

def findcost(distance):
        if distance is between 0 and 1
        set CostNew to 0
        return CostNew

else if distance is between 1 and 10
        distance1 equal to distance minus 1
        set distancepercurrency equal to 5.5
        CostNew equal to distance1 multiply by distancepercurrency
        return CostNew

else if distance is between 10 and 20
        distance2 equal to distance minus 10
        set distance1 to 9
        set distancepercurrency1 to 5.5
        set distancepercurrency2 to 6.5
        CostNew equal to distance1 multiply by distancepercurrency1 and plus with distance2
        multiply by distancepercurrency2
        return CostNew

else if distance is between 20 to 40:
        distance3 equal to distance minus 20
        set distance2 to 10
        set distance1 to 9
        set distancepercurrency1 to 5.5
        set distancepercurrency2 to 6.5
        set distancepercurrency3 to 7.5
        CostNew equal to distance1 multiply by distancepercurrency1 and plus with distance2
        multiply by distancepercurrency2 then plus with distance3 multiply by
        distancepercurrency3
        return CostNew
```

```
else if distance is between 40 to 60
        set distance4 equal to distance minus 40
        set distance3 to 20
        set distance2 to 10
        set distance1 to 9
        set distancepercurrency1 to 5.5
        set distancepercurrency2 to 6.5
        set distancepercurrency3 to 7.5
        set distancepercurrency4 to 8
        CostNew equal to distance1 multiply by distancepercurrency1 and plus with distance2
        multiply by distancepercurrency2 and plus with distance3 multiply by
        distancepercurrency3 then plus with distance4 multiply by distancepercurrency4
        return CostNew
else if distance is between 60 to 80
        set distance5 equal to distance minus 60
        set distance4 to 20
        set distance3 to 20
        set distance2 to 10
        set distance1 to 9
        set distancepercurrency1 to 5.5
        set distancepercurrency2 to 6.5
        set distancepercurrency3 to 7.5
        set distancepercurrency4 to 8
        set distancepercurrency5 to 9
        CostNew equal to distance multiply by distancepercurrency1 and plus with distance2
        multiply by distancepercurrency2 and plus with distance3 multiply by
        distancepercurrency3 and plus with distance4 multiply by distancepercurrency4 then
        plus with distance5 multiply by distancepercurrency5
        return CostNew
else if distance more than 80
        set distance6 equal to distance minus 80
        set distance5 to 20, set distance4 to 20, set distance3 to 20, distance2 to 10
        and set distance1 to 9
        set distancepercurrency1 to 5.5 and set distancepercurrency2 to 6.5
        set distancepercurrency3 to 7.5 and set distancepercurrency4 = 8
        set distancepercurrency5 to 9 and set distancepercurrency6 to 10.5
        CostNew = distance1 multiply distancepercurrency1 and plus distance2 multiply by
        distancepercurrency2 and plus distance3 multiply by distancepercurrency3 and plus
        distance4 multiply by distancepercurrency4 and plus distance5 multiply by
        distancepercurrency5 then plus distance6 multiply by distancepercurrency6
        return CostNew
```

Appendix B

Profit Analysis Result and Additional Data

Before we start to build our final model, we have test and explore previous candidate models such as random forest, gradient boost regression tree, and decision tree. The remaining result are as following section.

B.1 Root Mean Square Error(RMSE) Comparison

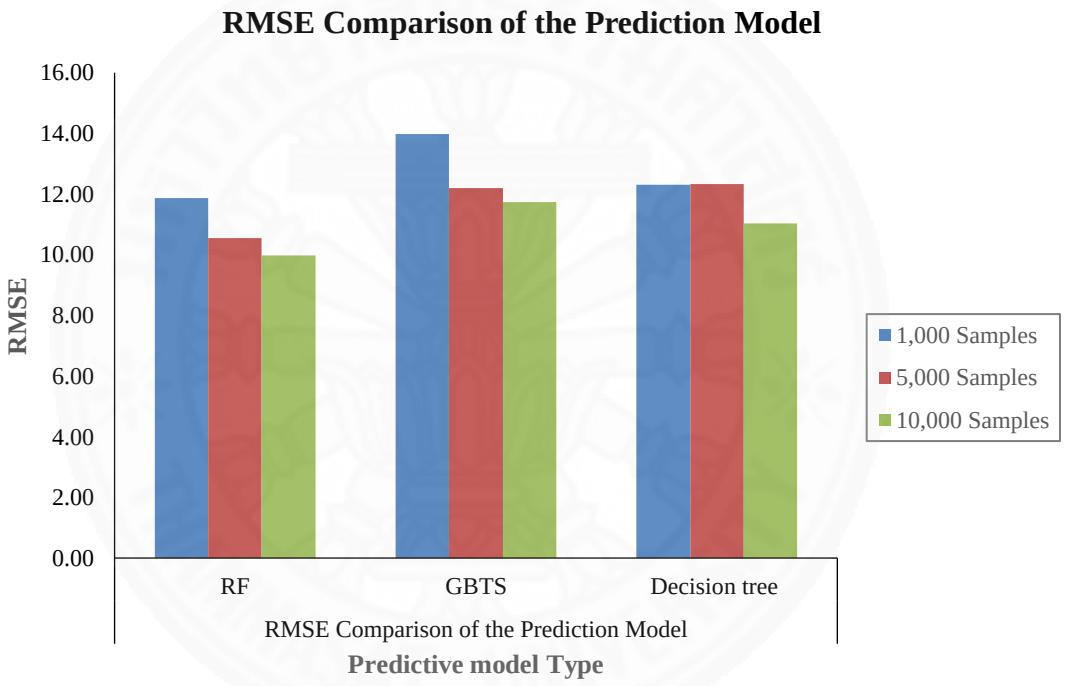

Figure B.1: Root Mean Square Error Comparison

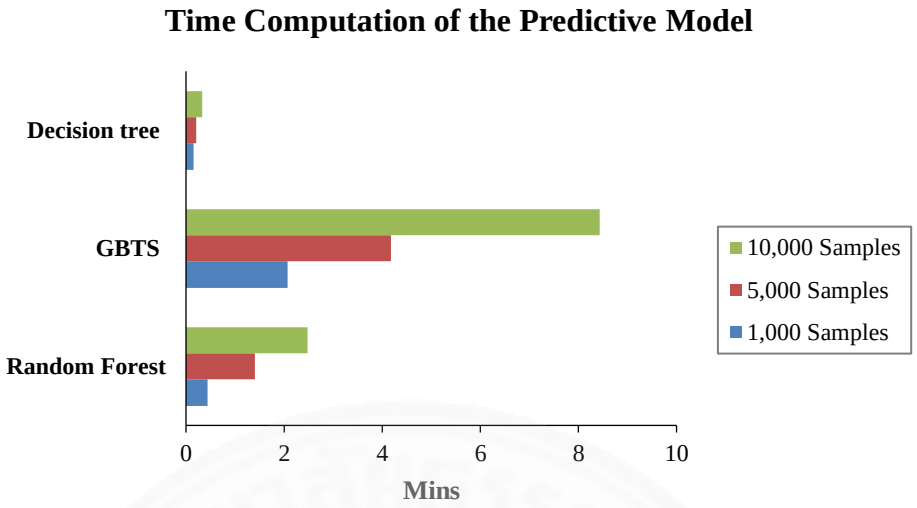

Figure B.2: Time Computation of the Predictive Model

From this test, it shown that if we make good tuning the random forest will reduce error easier than the remaining two model and give optimal execution time with high accuracy.

B.2 Net Profit per Distance Visualization

We visualize taxi's trip worthiness level in every 3 hours as following figures

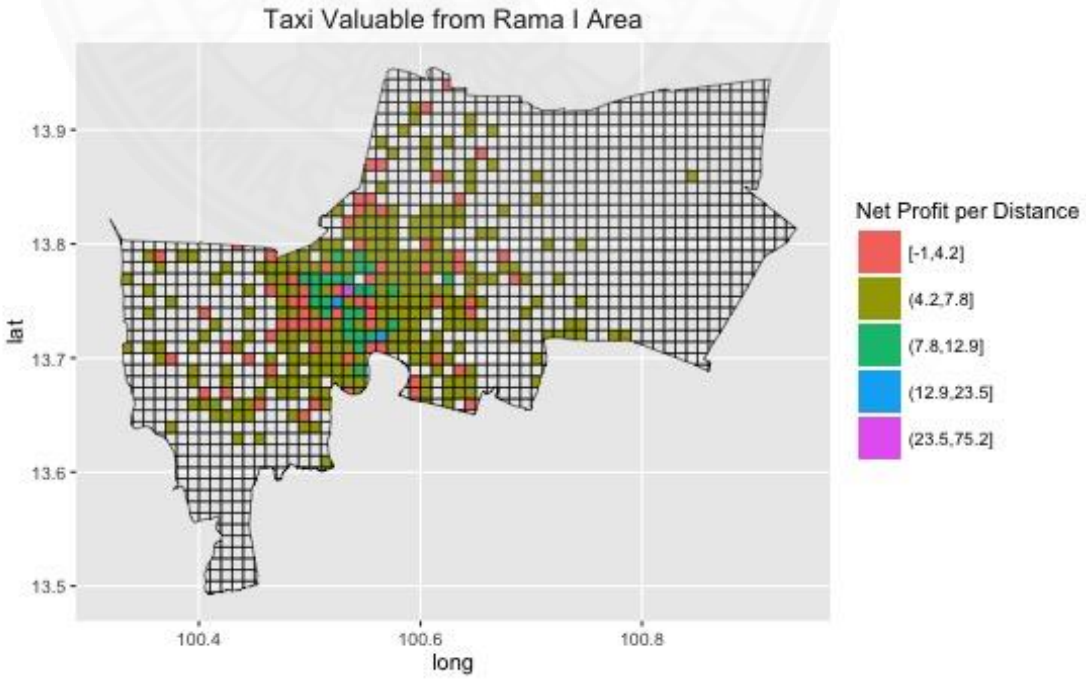

Figure B.3: Worthiness Value from Rama I (0.00 Hr.)

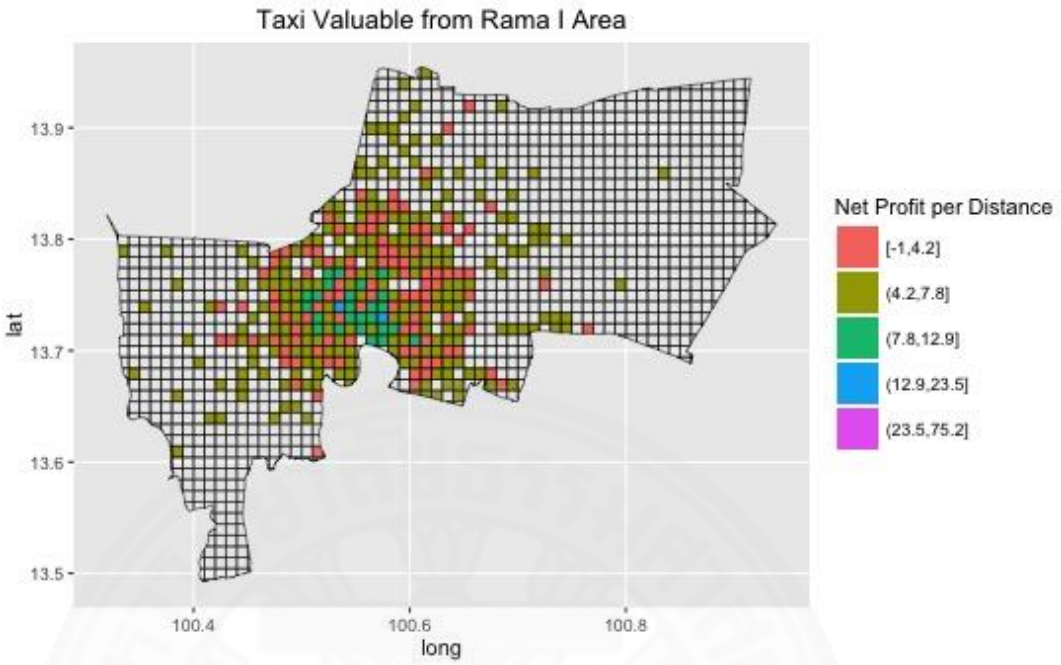

Figure B.4: Worthiness Value from Rama I (3.00 Hr.)

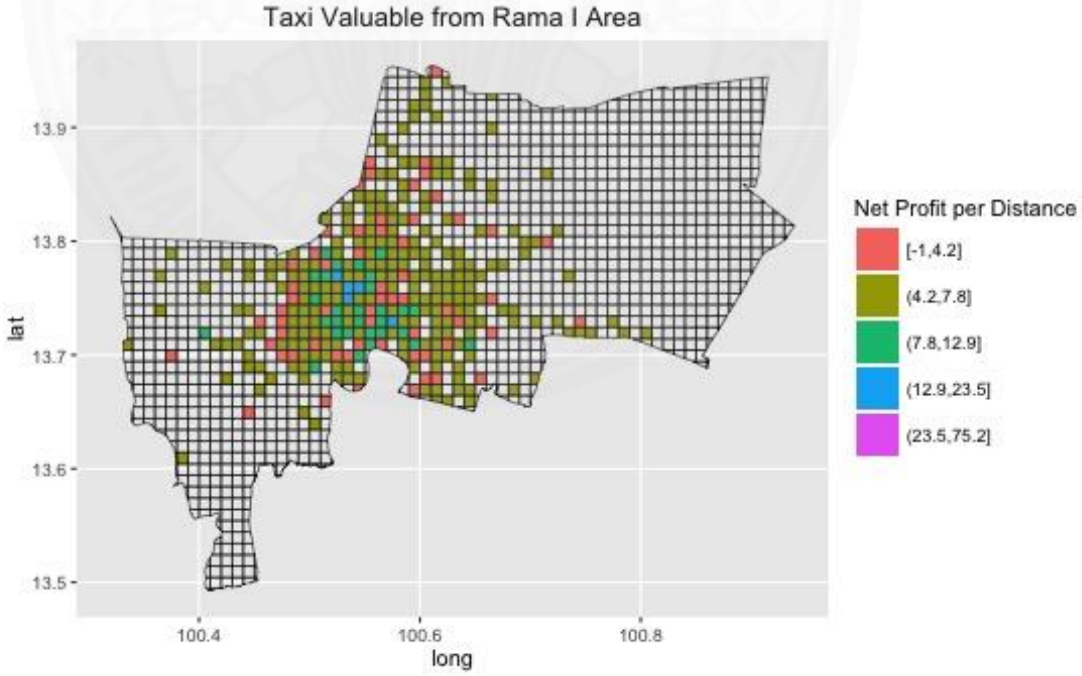

Figure B.5: Worthiness Value from Rama I (6.00 Hr.)

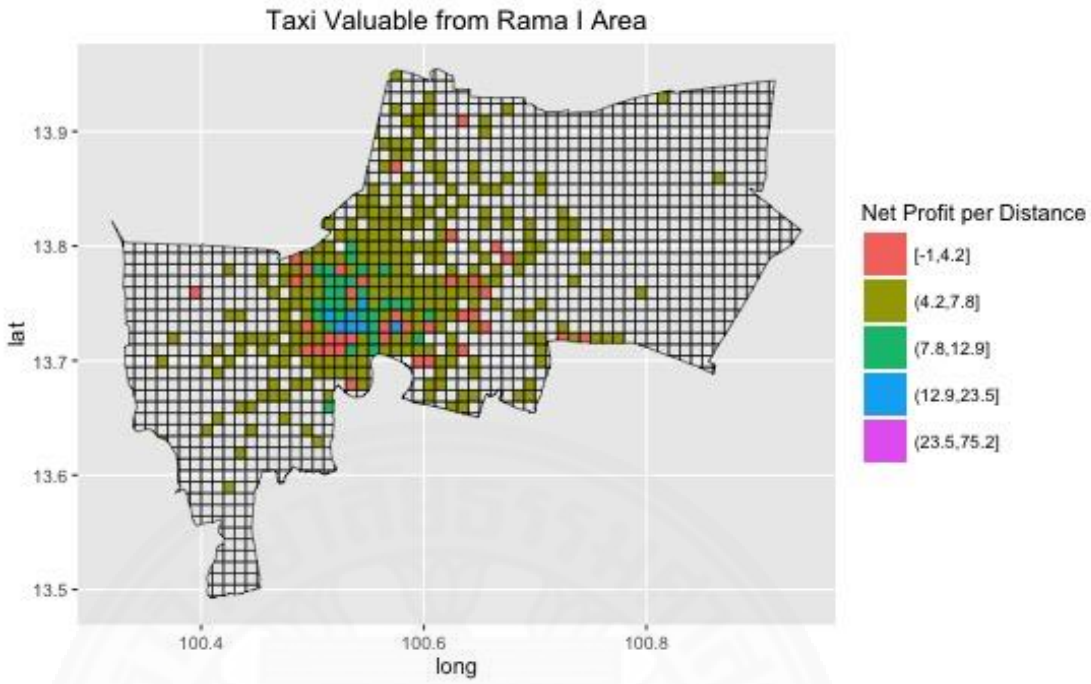

Figure B.6: Worthiness Value from Rama I (9.00 Hr.)

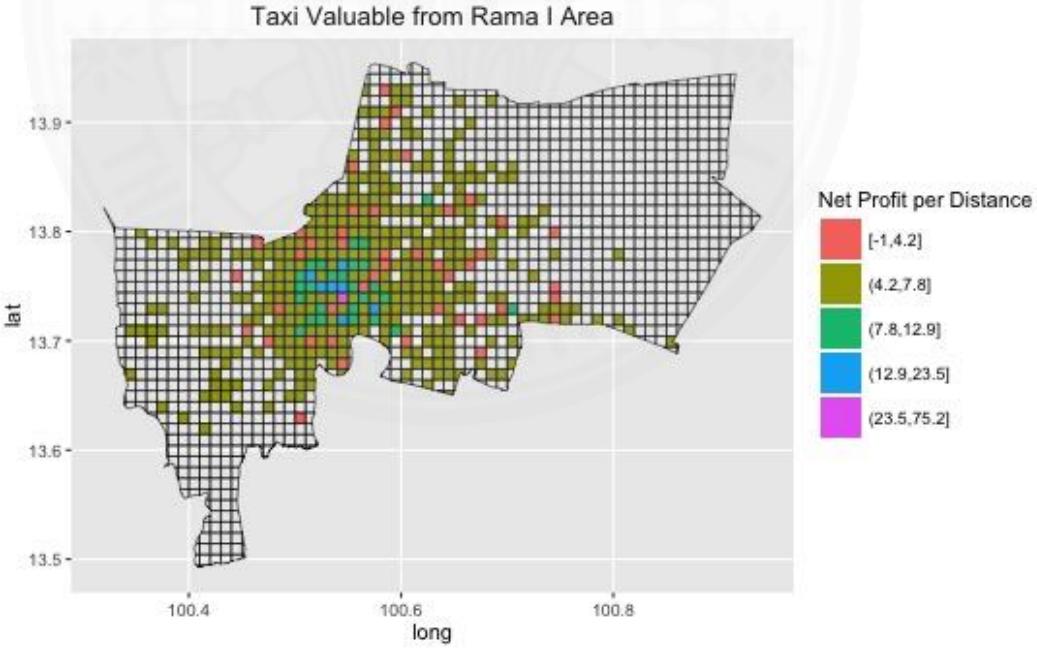

Figure B.7: Worthiness Value from Rama I (12.00 Hr.)

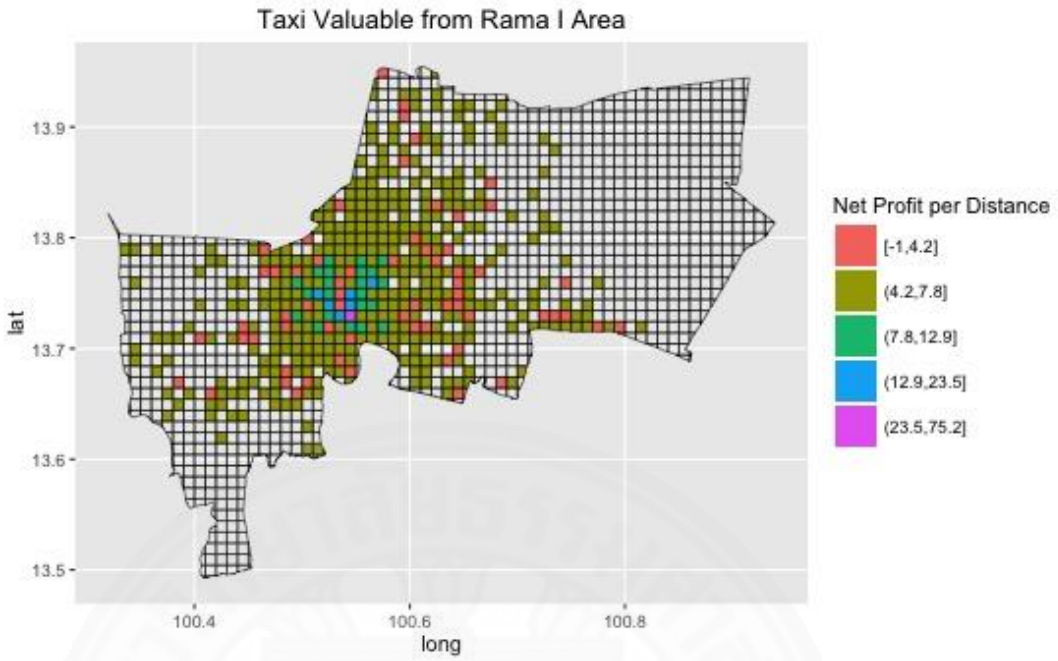

Figure B.8: Worthiness Value from Rama I (15.00 Hr.)

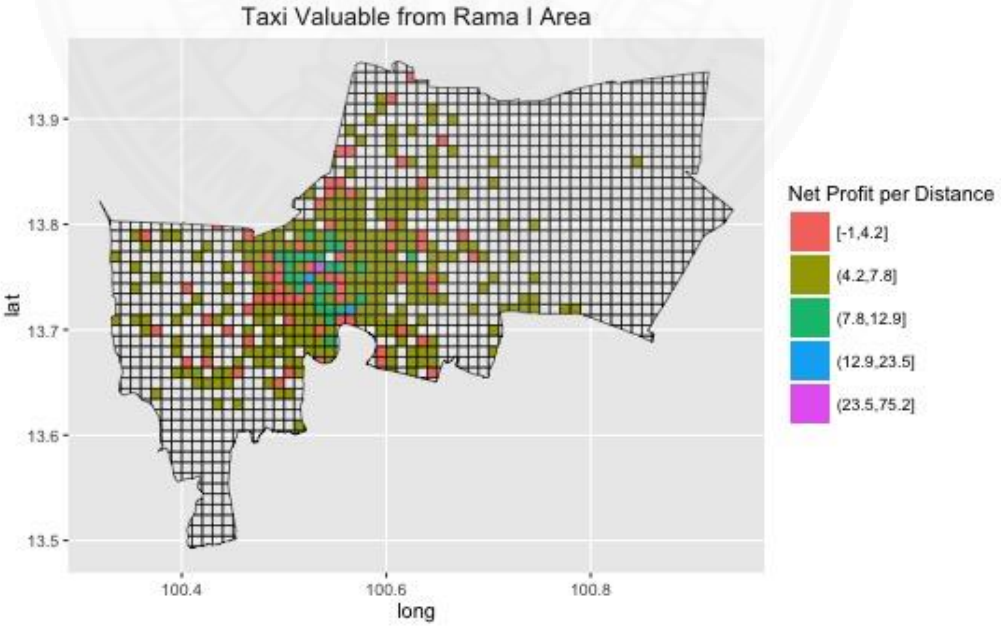

Figure B.9: Worthiness Value from Rama I (18.00 Hr.)

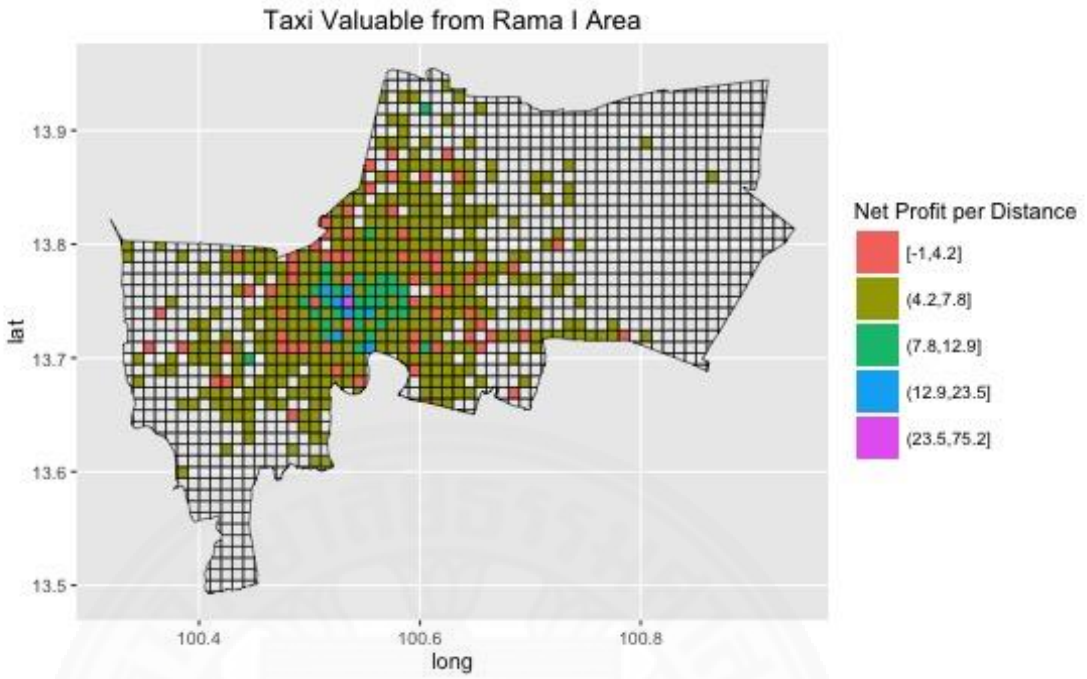

Figure B.10: Worthiness Value from Rama I (21.00 Hr.)

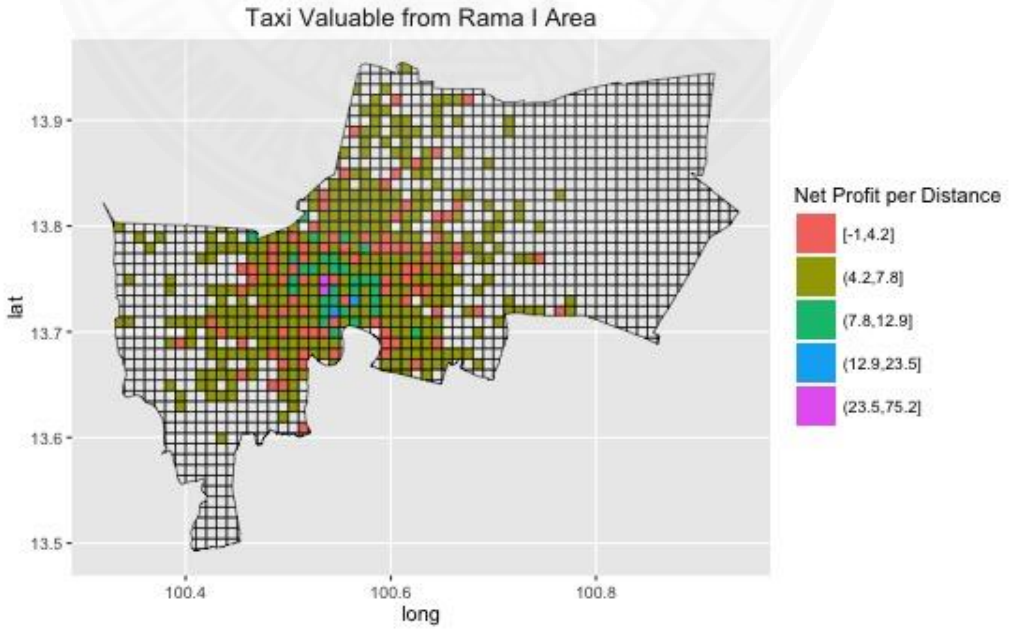

Figure B.11: Worthiness Value from Rama I (23.00 Hr.)

B.3 Taxi Trip Worthiness level over 1 million Trips

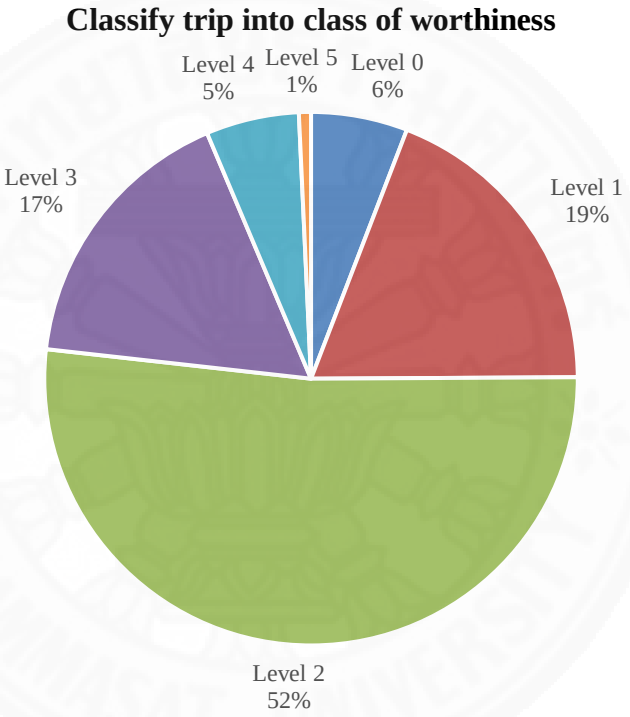

Figure B.12: Taxi Trip Worthiness Level Classification Result

B.4 Additional Taxi Trip Basic Statistic

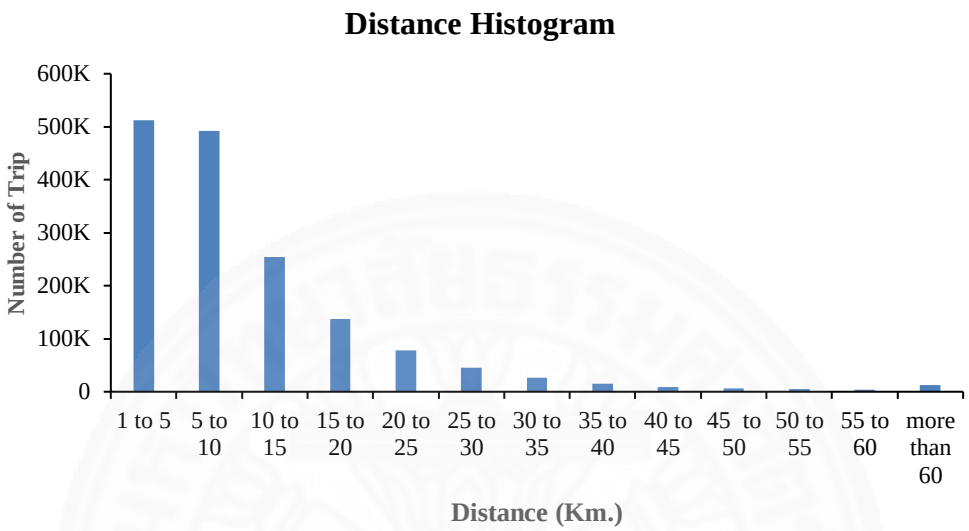

Figure B.13: Distance Histogram (Range every 5 Km.)

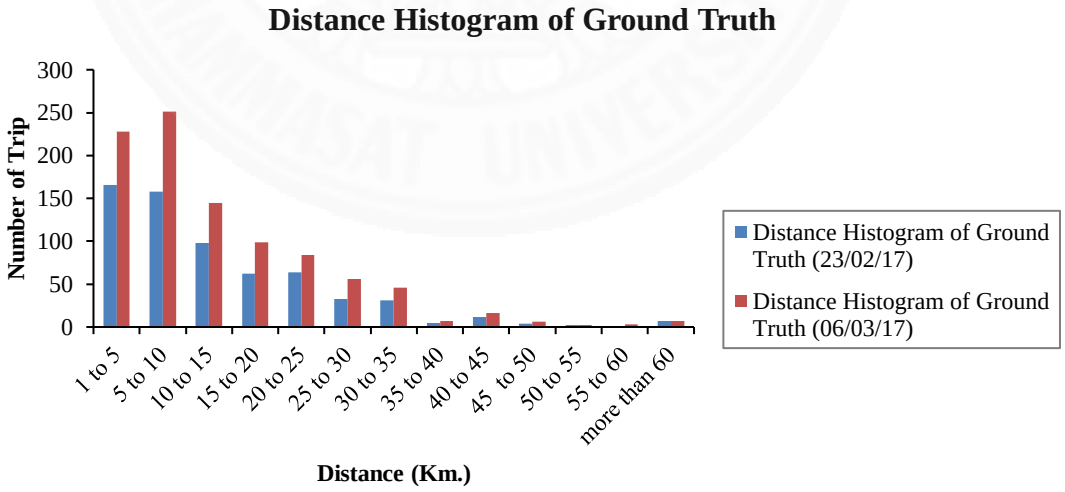

Figure B.14: Distance Histogram of the ground truth data

B.5 Prediction Result of Profit Model (06/03/17)

Table B.1: Initial prediction model result

	Actual Profit	Predicted Profit	RMSE (Baht)	Accuracy (%)	Error (%)	R-Squared
Model by Simulation (Calculation from algorithm):5 Tree Max depth 30, max bin 32						
70:30 Random Split	120.02	129.72	9.70	0.97	8.09%	91.91%
Test with all data	120.02	135.17	15.15	0.94	12.62%	87.38%
Model by Simulation (Calculation from algorithm):10 Tree Max depth 30, max bin 32						
70:30 Random Split	120.02	129.19	9.17	0.97	7.64%	92.36%
Test with all data	120.02	133.04	13.02	0.95	10.85%	89.15%
Model with Algorithm Result with 10,000 of training data						
70:30 Random Split	120.02	128.97	8.95	0.97	7.46%	92.54%
Test with all data	120.02	129.48	9.46	0.98	7.88%	92.12%

Appendix C

List of Publications

C.1 International Conference

1. Thananut Phiboonbanakit and Teerayut Horanont, Who will get benefit from the new taxi fare rate? Discerning the real driving from Taxi GPS Data, Proceedings of the International Conference of Information and Communication Technology for Embedded Systems (ICICTES 2016), March 20-22, 2016, Thailand.
2. Thananut Phiboonbanakit and Teerayut Horanont, Understand Trend of Taxi Usage and its Fare Rate from Large-Scale Analysis of Trajectory Data, Proceedings of the 4th International Symposium on Fundamental and Applied Sciences (ISFAS) March 29-31, 2016, Kyoto, Japan.
3. Thananut Phiboonbanakit and Teerayut Horanont, How does taxi driver behavior impact their profit? Discerning the real driving from large scale GPS traces, Proceedings of The 5th International Workshop on Pervasive Urban Applications, In conjunction with ACM UbiComp 2016, 13 September 2016, Heidelberg, Germany.