Journal of Big Data

RESEARCH Open Access

Analyzing Bangkok city taxi ride: reforming fares for profit sustainability using big data driven model

Thananut Phiboonbanakit[1,2*] and Teerayut Horanont[1]

*Correspondence: d6022300021@g.siit.tu.ac.th
[1] School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
Full list of author information is available at the end of the article

Abstract

With the trend toward the use of large-scale vehicle probe data, an urban-scale analysis can now provide useful information for taxi drivers and passengers. Unfortunately, traffic congestion has become a critical problem in urban cities. Road traffic congestion reduces productivity in transportation services, and the daily profit earned is consequently reduced. This is opposite to the cost of living, which is increasing rapidly. Therefore, these issues are causing difficulties in all occupations in terms of managing daily expenses, particularly for taxi drivers. The taxi driving is classified as low income compared to other occupations. Such facts are a symbol of economic inefficiency. To this end, this study aims to assist taxi agencies and the government in improving taxi driver profits in Bangkok using large-scale data. To deal with these large-scale data, we propose a big data-driven model. With this model, we first calculate costs using a cost–distance algorithm and trip reconstruction. The data are then modeled to understand distance-based profits with respect to the departure time and traffic conditions. Finally, several cost predictive models using machine learning are evaluated using the ground truth from 50 taxis for a 1-month period. The experiment results show that more frequent trips over a short distance yield higher profits than long-distance trips. Finally, a solution to improve taxi driver profits is determined. We also compare the advantages and disadvantages of a unified solution.

Keywords: Global positioning system, Driving behavior, Profit assessment, Spatial–temporal analysis, Big data

Introduction

Numerous innovative mobile services have been introduced into society and have caused many changes in our daily lives. People can use various devices to serve their needs and interact with such devices as personal assistants. On the street, in-vehicle tracking is widely used to collect massive amounts of trajectory data, and with today's rapid progress in big data analysis, such mobility data can be mined to evaluate traffic congestion, travel speed, and trip time in urban areas. The results provide us with new insights into understanding traffic phenomena and can help mitigate various driving problems.

Bangkok is one of the world's most congested cities, making it a challenge to get around during peak rush hours. Traffic congestion has been one of the most debatable issues for many years. The study in [1] shows that in Thailand, public and private vehicles spend additional time on the road and pay up to 16% of additional fuel costs for transportation. Unfortunately, road traffic congestion reduces the productivity of transportation services. Consequently, the profit earned daily is reduced. However, this is opposite to the cost of living, which has been increasing rapidly in recent years.

The easiest and most comfortable way to get around the city, if not always the quickest, is by taxi. Thailand is one of the world's major cities with the cheapest taxi fares [2]. For one taxi ride at a distance of 5 km, the fare is only 57 THB or 1.5 EUR. The initial fare for a taxi ride is also cheaper than the cost of living. For instance, a taxi fare is cheaper than the price for a single bottle of beer [3].

Bangkok has hundreds of thousands of taxicabs and finding one at any time is never a problem. However, getting a taxicab in Bangkok can sometimes be challenging. In many cases, the driver will not accept rides to areas with heavy traffic. As mentioned earlier, road traffic congestion reduces the service productivity and reduces daily profits. This issue results in drivers avoiding trips toward congestion areas. In addition, the driver might be at the end of the driving shift (e.g., day or night shift), and intended destination might be in the opposite direction as the vehicle return location. Therefore, taxi drivers most commonly refuse to pick up a customer when the passenger's destination differs from the driver's preferred route. Indeed, if the passenger's route needs to go through heavy traffic, it will reduce the driver's profit.

Taxicabs play an essential role in city transportation networks. They have been shown to be an important transportation mode in most large cities globally. However, in many countries, economic issues as well as the presence of alternative modes of transportation have a negative impact on taxi driver incomes. This issue has been widely discussed in many different sectors, although no real solution has been reached. This problem leads to taxi drivers frequently rejecting passengers because they would like to service customers that will gain them more profit. A news article in 2015 stated that "Bangkok is a city infamous for its taxi service, with drivers often rejecting passengers, demonstrating poor road manners and charging overpriced fares."

It is not only local citizens who face this problem but also travelers visiting Thailand. In our study, we further conducted a survey and interviewed taxi drivers in Bangkok. The results show that 62% of taxi drivers have had experiences refusing passengers, the potential reason being that the passenger's desired route would gain them less profit. The issues mentioned above motivated us to study and propose a sustainable solution to this problem.

The key point of interest is the data applied to the system (e.g., data from a GPS tracker installed in the vehicle), which are fast, massive, and are of significant variety. Such data have been called big data.

As a benefit of big data, they contain vital information that can extract insight into mobility patterns. Such information can be used to help taxi drivers earn more money. For instance, they reveal the mobility patterns of taxi trajectories. We can therefore know how the driver operates the vehicle and the amount of profit earned for each ride. Furthermore, big data can be used to recommend what not to do to avoid a reduction in income.

However, it is not easy to extract such information from big data. Occasionally, such data have noise and various outliers. To this end, the present study applied a big-data-driven model for dealing with such massive data amounts. We aim to use this information to assist taxi agencies and the government in improving taxi driver profits in Bangkok. A solution to the above issues was also found through this study.

The remainder of this paper is organized as follows. "Related studies" section focuses on a survey of methodologies for dealing with large-scale data. "Problem definition" section provides a detailed statement of the problem, its significance, and the motivation behind this research. "Methodology" section presents the methods used to conduct the data analysis process, as well as the experiment conducted using a data analysis and suggestions for areas of improvement. The proposed model is also described. The next section presents the experimental results. "Discussion" section describes the factors involved in a taxi profit analysis, as well as a comparative experiment conducted. Finally, our findings are summarized, and some concluding remarks and suggestions for potential improvements to the present study are provided.

Related studies

In recent studies, many researchers have tried to find a way to address this issue. Such studies can be divided into various perspectives including spatial and temporal, recommendation systems, demand prediction, algorithms, and application developments. Large-scale data can be used in an analysis of the spatial and temporal locations; for example, [4] mined interesting locations and travel sequences in specific geospatial regions using GPS trajectories on a tree-based hierarchical graph to find the neutral relationship of the visitors in the region. In addition, [5] used taxicab trajectory data to discover popular areas.

Some studies have used data to develop a recommendation system for improving driver profits; for example, [6] proposed a taxi recommendation system for determining the next cruising location using an L–L graph model. Qu et al. [7] developed a cost-effective recommender system for taxi drivers, and Kamimura et al. [8] presented a recommender system, called D-Taxi, that informs taxi drivers where to find the next passenger using the latest pickup and drop-off data. Moreover, Ding et al. [9] defined a new method called global-optimal trajectory retrieving (GOTR) to find a route with high profit and increase the chance of receiving the next customer. Zhang et al. [10] proposed a pick-up recommendation method for taxi drivers based on spatio–temporal clustering. Yuan et al. [11] presented a recommender system for taxi drivers and passengers seeking a taxicab using the knowledge of (i) passenger mobility patterns and (ii) taxi driver pickup behaviors. Finally, Zhang et al. [12] proposed a cruising system, pCruise, for taxi drivers to maximize their profits by finding the optimal route to pick up a passenger.

Large-scale data can also be used for demand prediction within certain periods of time. For example, Qi et al. [13], presented a method to predict the waiting time for a passenger at a given time and spot from historical taxicab trajectories. Zhang and Haghani [14] used a gradient-boosted tree to predict and improve the traveling time. Moreira et al. [15] used taxi data to predict demand by using streaming data, and Yingjun et al. [16] predict the number of taxicabs using wavelet-based neural networks. Finally, Zhou et al. [17] predict bus passenger demand based on mobile device usage.

In additional to such studies, the benefit of large-scale data can be applied. For example, [18] introduced an algorithm to calculate taxi fare rates from large-scale data. The main objective is to determine the impact on driver profits when applied a different fare within various distance range. Bai and Wang [19] developed a taxicab routing and fare rate estimation by integrating the calculation algorithm into a mobile application, and Egan and Jakob [20] created a marketing design for on-demand transport services that considers market mechanisms in the design.

Fortunately, in recent years, improving taxis services has become a popular topic attracting many researchers in the transportation research field. Zhang et al. [21] proposed a queuing network approach to improve the customer's waiting and searching time. This model also considers traffic congestion on urban road links. Jayasooriya and Bandara [1] presented a methodology for measuring the economic costs of traffic congestion. The authors discovered that road traffic congestion reduces the workforce productivity for both private and public motorized transport. Regrettably, the cause of high traffic congestion originates from bus transportation and car/van transportation.

Further, Wong et al. [22] proposed a methodology to improve taxi services. First, a survey of customers was conducted to rate the level of satisfaction regarding a taxi service. Next, an enhanced linear regression model was developed to identify the priority areas for improvements in the service quality of urban taxis. Finally, a six level of service (LOS) tool was applied to analyze the current level of a taxi service. Moreover, Čulík et al. [23] proposed a cost analysis to analyze the taxi service cost on a digital platform (e.g., Uber, Bolt, and Taxify). Fortunately, their research is similar to this study, in that a cost analysis was conducted.

However, the critical difference is that [23] did not consider the traffic situation for the model analysis. The model seems to be static and is not driven by data. For [22], the approach is based on a paper-based survey. Their findings are impressive. However, we are currently in the era of a data-driven approaches, which are more feasible in the analysis stage than a regular survey. Unfortunately, using only a paper-based survey increases the time for data collection and analysis. Therefore, these issues, including road traffic congestion impacting the economic cost mentioned earlier, has motivated us to fill in the gaps in the studies by [22, 23].

In addition, the cost measuring methods presented in [1] are also adopted to compute the route cost of a taxi trajectory.

Problem definition

Road traffic congestion is a critical problem in urban cities. This congestion is a common issue whose resolution is debatable. The authors of [1] mentioned that road traffic congestion interrupts and reduces productivity in transportation services. People in urban areas spend more time on the road and are required to pay for the extra transportation cost. These issues have a wide impact on all occupations that apply transportation, particularly taxi drivers.

Unfortunately, taxis service demand has changed over time. This has caused difficulties for taxi drivers to maintain their profits based on the road traffic congestion. To this end, it is necessary to discover the cause of such congestion and suggest a feasible solution to better assist taxi drivers.

However, the data obtained are highly dimensional, making it more challenging to analyze than single-dimensional data. Each dimension consists of vital information. Therefore, it is necessary to develop a new methodology to handle such data.

If a traditional approach is applied, it will result in drivers potentially losing opportunities to serve customers, with a decrease in overall profit, consequently impacting the overall economy.

To overcome these issues, a new big-data driven-model is proposed. This model is designed to handle large-scale data. After the data are transformed, a distance-cost based model is applied to extract the trajectory information. The information is then put into the predictive model using machine learning (ML).

As a result, the accuracy of the prediction results indicates whether the data patterns have been accurately extracted. Finally, the extracted information is then used to suggest an optimal fare addition for taxi drivers in finding and picking up customers.

Methodology

There are approximately 140,000 official licensed taxis in Bangkok. We collected the GPS probe data from 10,000 taxis with 5-s intervals for a 5-month period in 2016 (January through May 2016). Approximately 5000 taxis are active at the same time. An analysis of this large-scale mobility data allows us to achieve insight and understand the potential impact of profits of taxi drivers based on their detour routes and traffic profiles using big-data-driven models. The results obtained in this research can provide better information regarding taxi fares and indicate better solutions to the taxi routing problem.

We assume that the profits of taxi drivers will vary depending on the distance and traffic congestion during their trips. The current taxi fares in Bangkok were calculated. The fixed starting fare for a distance of 0–1 km is 35 THB. For a running distance of 1–10 km, the taxi meter rate is set to 5.5 THB per km. The Bangkok taxi fare changes by 2 THB; therefore, the meter will go up every 0.36–0.37 km at this rate. In other words, the longer the distance of the taxi trip, the higher the taxi meter rate.

However, if the distance is over 10–20 km, the rate is increased to 6.5 THB per km. The meter will go up by 2 THB every 0.30–0.31 km. Further, from 20 to 40 km, the rate is increased to 7.5 THB per km and continues increasing until reaching 80 km, where it maxes out at 10.5 THB per km. In addition, when a taxi is in a traffic jam (i.e., moving slower than 6 km/h), an extra charge of 2 THB per minute is added to the current fare. For more information regarding taxi fares, please refer to [24].

Data cleansing and preprocessing

We first obtained taxi GPS data for more than 5000 taxi vehicles in Bangkok from January to May 2016. These data are considered big mobility data, with the size of approximately 5 terabytes. Therefore, we believe that the data covers all trends regarding the mobility of taxis and are ready for analysis. However, the data contain many outliers and errors, and we first needed to explore and clean the data for their removal. Therefore, we first created a $1\,\text{km} \times 1\,\text{km}$ geospatial grid and used a spatial intersection technique to divide the data into a small sub-area.

The origin–destination (OD) matrix is created and referenced by a 1 km^2 area. The GPS data that are not located in the Bangkok area were removed. In addition, we cleaned up existing errors and outliers by applying statistical rules such as the speed, distance, and total amount of GPS point (GPS count) per trip. After the data went through the cleansing process, the grid windows are used to mapping the data, as shown in Fig. 1.

As presented in Fig. 1, the grid windows were created from the spatial polygon and divided into 100 × 100 squares. In addition, each 100 m × 100 m grid is assigned a unique ID. During the next step, we took a road network from OpenStreetMap [25] and the intersection technique [26] are then applied to the spatial polygon and the road network. As a result, the grid intersected with the road network is returned along with the unique grid ID.

After we obtained a grid, we then matched the GPS data point to the grid using the intersection technique. Therefore, each GPS data point is assigned along with the grid's unique ID, as shown in Fig. 1. This unique ID is further used when we reference the location of the trajectory or when we want to remove any outliners in the GPS data points that occur far from the road (e.g., do not have a unique ID). This technique aims to enhance the quality of the trajectory. This is vital because the lower the error, the more accurate the distance and travel information that are returned.

Second, we explored the data components such as the speed, data source, meter, timestamp, and International Mobile Equipment Identity (IMEI). We removed the data from unknown data sources that were not related to this study.

The proposed model

After obtaining the preprocessed data, we then constructed a model using a data-driven framework, which is called a big data-driven model. The model is presented in Fig. 2.

From Fig. 2, the proposed model consists of six components as follows:

- Data,
- Cost–distance algorithm,
- Trip basic statistic report,
- Data analysis,
- Profit prediction and result evaluation,
- Profit recommendation.

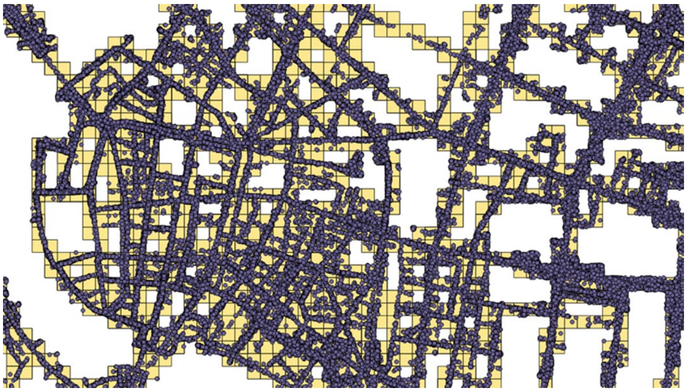

Fig. 1 Road segments mapped to a 100-m grid

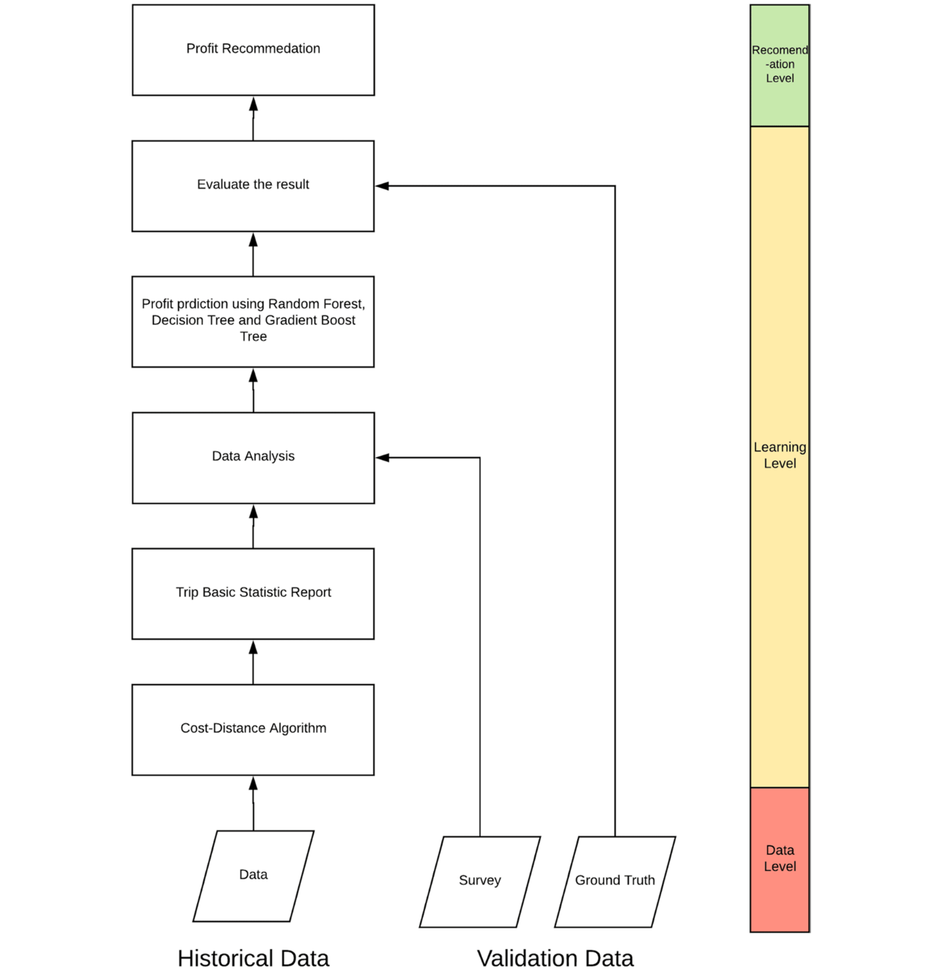

Fig. 2 Data-driven framework for analyze taxi probe data

In the following sections, we then presented the methodology behind these components.

Data

The data that are processed in the data pre-processing stage are now ready to be input into the model. The data structure is shown in Table 1.

Cost–distance algorithm

We developed a cost–distance algorithm that improved from [18, 27]. This algorithm aims to determine the distance, average speed, total trip time, and trip cost. In addition, this variable is used to analyze the impact on a taxi driver's revenue (net profit) in the following sections. The algorithm used to process the data is presented in Algorithm (1). Please note that the code for this algorithm will be available at [28].

Table 1 Structure of the taxi GPS data collected from real driving vehicles in Bangkok

Field	Variable	Description
IMEI	"10011304"	Taxi identification
Latitude	13.73522	Degree
Longitude	100.58979	Degree
Speed	30.0	km/h
Direction	16	Degree
Error	0.0	Data error status
Acceleration	0	Paddle press status
Meter	1	Taxi's vehicle vacant status (0 mean vacant and 1 mean has passenger)
Time-stamp	2016-01-14 10:39:40	Date and time of GPS location

In this study, a cost–distance algorithm was developed and used for transforming the structure data of GPS data points into the origin–destination (OD) trip information. The data are stored and ordered using IMEI, a timestamp in a hive table of a Hadoop cluster.

The algorithm took the streaming GPS point, speed, meter status, and timestamp to begin the computation process. Please note that the algorithm takes one record at a time for computation. First, the algorithm checks for the meter status. If the meter status is equal to "1," then the IMEI, date–time, speed, and coordinates at the origin are stored into a temporary variable.

Second, the next GPS record is fetched and input into the algorithm. Similar to the previous step, the algorithm checks for the meter status. Suppose that the meter status is equal to 1 and belongs to the same IMEI that is stored in memory. In this case, the cumulative distance, travel time, and traffic congestion (e.g., when the vehicle is moving slower than 6 km/h.) are computed. The distance is computed from the cumulative distance between the two points of the origin and current coordinates.

Similarly, for the travel time and traffic congestion, there is a time difference between each point. Finally, the IMEI, date–time, and speed are updated into a temporary variable. In addition, the profit earned during the trip is also computed at this stage. The computed profit follows the guideline in [24], and remains at this stage until a meter status equal to "zero" is fetched into the algorithm.

Finally, the final profit at the destination is computed from the cumulative distance and traffic congestion time obtained earlier. In addition, the average speed was computed at this stage. As a result, the profit at the current trajectory is returned for storing into a hive table and preparing for further analysis.

This algorithm continues the procedure mentioned above until all GPS records are computed. In summary, the result consists of the distance, average speed, total trip time, and profit for each trip. The result is stored in a hive table, which is arranged and ordered by date and time.

For the specifications of the computations, the algorithm was run on Apache Hive inside the Hadoop cluster. The cluster machine consists of eight machines. Each machine consists of an Intel Xeon CPU, 1 TB of storage, and 16 GB of RAM. Therefore, the time

Algorithm 1: Cost-distance algorithm

```
Input  : Dataset $D = \{\text{distance}, total_{triptime}, \ldots F_n\} \in \Re$
Output: $IMEI, lat, lon, olat, olon, dlat, dlon, distance, total_{triptime},$
        $traffic_{delay}, profit, meter, dt$
for $i = 0 : D_{size}$ do
    if ($p_{meter}$ == "None" or 0) and ($meter = 1$) then
        % This part stores information on the trip's origin
        $p_{meter} \leftarrow meter$
        $p_{grid} \leftarrow grid$
        $p_{dt} \leftarrow dt$
        $p_{lat} \leftarrow lat$
        $o_{lat} \leftarrow lat$
        $p_{lon} \leftarrow lon$
        $o_{lon} \leftarrow lon$
        $i \leftarrow i + 1$
    else if ($p_{meter} = 1$) and ($meter = 1$) then
        % Compute cumulative distance, travel time and traffic congestion of the
        current trajectory
        $distance \leftarrow distance + \sqrt{(lat - p_{lat})^2 + (lon - p_{lon})^2}$
        $total_{triptime} \leftarrow total_{triptime} + \text{time}(Minonly[t] + Minonly[t+1])$
        if $Speed < 6$ then
            % Compute duration of traffic congestion when vehicle speed is less than 6
            km/h.
            $traffic_{delay} \leftarrow Traffic_{delay} + \text{traffic}(Minonly[t], Minonly[t+1])$
            $Speed \leftarrow push(Speed)$
        end if
        $Profit \leftarrow Profit + profit(distance, traffic_{delay})$ % Compute profit
        $p_{meter} \leftarrow meter$
        $p_{grid} \leftarrow grid$
        $p_{dt} \leftarrow dt$
        $p_{lat} \leftarrow lat$
        $p_{lon} \leftarrow lon$ %store current variables into temporary variables
        $i \leftarrow i + 1$
    else if ($p_{meter} = 1$) and ($meter = 0$) then
        $distance \leftarrow distance + \sqrt{(lat - p_{lat})^2 + (lon - p_{lon})^2}$
        $total_{triptime} \leftarrow total_{triptime} + \text{time}(Minonly[t] + Minonly[t+1])$
        $Profit \leftarrow Profit + profit(distance, traffic_{delay})$ % Compute profit after
        trip end
        $traffic_{delay} \leftarrow Traffic_{delay} + \text{traffic}(Minonly[t], Minonly[t+1])$
        $d_{lat} \leftarrow lat$
        $d_{lon} \leftarrow lon$
        $traffic_{delay} \leftarrow mean(speed)$
        $i \leftarrow i + 1$
    end if
end for
```

consumed for data processing is approximately 1.25–2.00 h, depending on the remaining machine resources.

Please note that we used Apache Hive and Spark for data computation. These two frameworks are conducted on different tasks. For instance, Apache Hive deals with data processing tasks. By contrast, Apache Spark conducts the data analysis part, such as the application of a predictive model and analytics. We understand that Apache Hive is slow. Therefore, it was used only for dealing with the data processing tasks.

We chose Apache hive for running this algorithm is as follows:

1. The algorithm deals with data processing, which extracted the taxi's trajectory statistics from the structure data.
2. Apache hive is a pure data warehouse and has its own SQL interface operating in Hadoop, the so-called "HiveSQL." Therefore, it is easy to develop a faster data warehouse framework.

In conclusion, the algorithm is presented in Algorithm 1 and its parameters are described as follows:

- The cost represents the cost of the current taxi trip, and the distance represent the trip's distance.
- The $total_{triptime}$ represents the time duration of the trip during a minute period.
- The $traffic_{delay}$ represents the duration in which the speed is less than 6 km/h (traffic jam) in minutes, and the meter represents the vehicle status (a 0 indicates no customer on board and a 1 indicates a customer is on board).
- The value of p_{meter} represents the previous meter status from the previous GPS data point.
- p_{grid}, p_{lat}, p_{lon}, and p_{dt} represent the previous grid, latitude, longitude, and timestamp from the previous GPS data point.

Trip basic statistic report

In this section, the basic statistic returned from the cost distance is presented. It shows patterns and insight on how taxi services are operated in Bangkok. Furthermore, a report consists of the basic statistics of an active vehicle ordered by time, vehicle speed, distance, and travel time.

We then explored the information; Fig. 3 shows all taxi trips categorized by date. It is obvious that a sudden decrease in active vehicles on January 1–3, 2016, is caused by the New Year's festival. This is because many drivers are absent. In addition, during April 12–16, 2016, a sudden decrease occurred again. This was caused by the Thai New Year festival. Thai people normally return to their hometowns and visit their relatives during this period.

Furthermore, Figs. 4, 5, 6 show the basic statistics on the speed, distance, and travel time of the taxi trips. These initial results are crucial because they allow recognizing and removing outliers from the data. As a result, we obtained high-quality results.

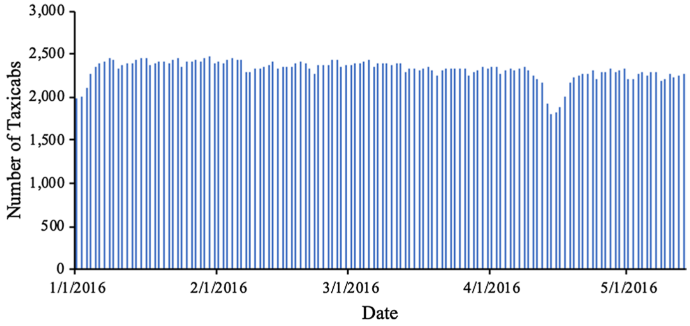

Fig. 3 Number of active taxicabs from January through May of 2016

Figure 4 shows that most of the taxi trips in Bangkok have an average speed of approximately 20–30 km/h because of the traffic conditions. Some of the trips consist of taxicabs searching for customers.

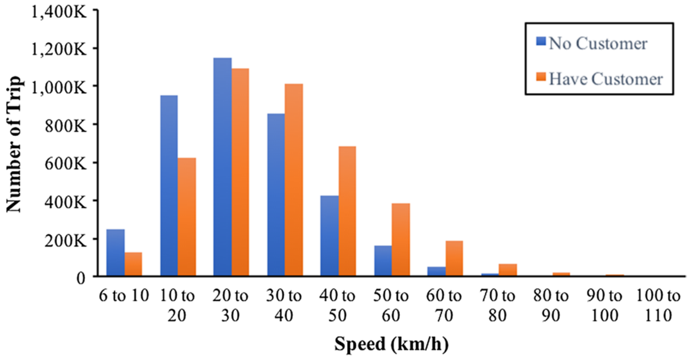

Fig. 4 Speed of taxicab trips

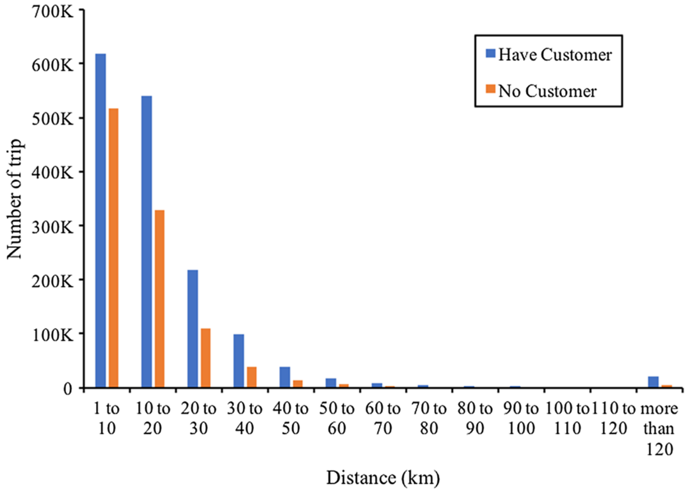

Fig. 5 Distance of taxicab trips

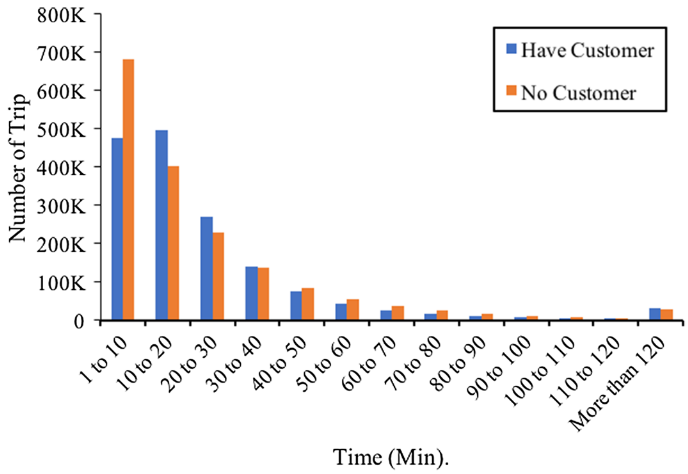

Fig. 6 Taxi trip duration

Figure 5 demonstrates that taxi trips in Bangkok are more likely to be of short distances than long distances. This result could point directly to taxi problems in Bangkok because some drivers are concerned with such distances, and aim to maximize their revenue with low operation costs.

Nevertheless, driving on long-distance trips increases their expenses. It also results in a low possibility of getting customers on the return trip.

Finally, the travel time was also determined, as shown in Fig. 6. It was shown that most of the trip of between 10 and 20 min long when they have customers on board. We can conclude that these two features depend on each other.

Taxi survey

In this study, the questionnaire survey and installation of an in-house mobile application aim to collect taxi service information from an experienced taxi driver in Bangkok. This information from the survey enables us to better understand the quality of the service, pricing, and target customers. It is also used to validate the predictive models shown in "Data analysis" section. These two data collection periods (ground truth) are 1 month in length.

The questionnaire survey asked the participants the following:

1. Gender.
2. Age.
3. How long have you worked as a taxi driver?
4. What type of vehicle do you drive?
5. Age of the vehicle.
6. Do you have regular maintenance of your vehicle?
7. What type of taxi service do you offer (public, private, or other)?
8. Have you ever rejected a ride to a passenger?
9. What is your average driving speed?
10. What is the most popular destination for your customers?

11. Which period is highly profitable?
12. Which period is difficult for you to get a passenger?
13. What is your most frequent trip length (short, medium or long)?
14. Which type of trip provides the highest profit?

Please note that this is an online questionnaire survey, and the full form is accessible through [29].

In addition to the questionnaire survey, an in-house mobile application was installed on the taxi driver model device, and was used to collect the driving trajectory and fare of each ride. The information from the mobile application is also used for model validation.

The details of the information collected from the mobile application are as follows:

1. Taxi fare for each ride. The fare is in THB.
2. Same features mentioned as mention in "Profit prediction and result evaluation" section.

Data analysis

In the previous section, statistics from various features were analyzed. Therefore, in this section, the equations for calculating the net profit and expense costs are introduced. The equations are calculated based on the trajectory distance, travel time, and traffic delay (i.e., vehicle moving at a speed of less than 6 km/h).

Finally, the output equation is the taxi driver's net profit for each trip. After the net profit is returned, we then analyze the net profit based on the area of the pickup, the distance of the trip, and the driving period.

Net profit and expense costs

The results obtained from the cost–distance algorithm were used to calculate the net profits (*N*). The net profit (*N*) was calculated as shown in Eq. (1).

$$N = Profit - \left(\frac{d \times ct}{md}\right) - (T \times S) - TF - \left(\frac{pd \times ct}{md}\right) - (PT \times S) - PTF, \tag{1}$$

where d represents the distance that can be traveled in the current time, T represents the total trip time, pd represents the distance traveled before the customer is picked up, and PT represents the total trip time before the customer is picked up. The labor and vehicle maintenance costs are not included in this equation. There are two main reasons for this. First, there are many types of vehicles that operate as taxis. Second, taxi drivers are freelancers and do not have any monthly fixed allowance from the company.

The overall details of this equation are divided into three parts. The first part denotes the profit that is collected from the passenger. The second part denotes when a taxi vehicle has a customer (passenger) on board. The third part denotes the cost at which the taxi cruises to find customers. After deducting this cost, it is the net profit where taxi drivers earn on each trip.

In this study, the Toyota Corolla Altis 1.6 CNG was used as the sample vehicle, which is commonly used as a taxi vehicle in Thailand. The vehicle specifications include a 55-L fuel tank and a 75-L natural gas for vehicle (NGV) tank. The fuel consumption was 12.19 km/L, which was obtained from the manufacturing documents. The fuel costs are 22.04 THB/L for fuel and 13.36 THB/L for NGV (as of May 5, 2016). The cost for a fulfilling vehicle fuel tank (*ct*) is given by Eq. (2).

$$ct = fuel_{tank} \times fuel_{price}. \tag{2}$$

The details of Eq. (2) are a multiplication of the vehicle capacity fuel tank by the fuel price. Therefore, the fuel cost is returned. Moreover, *md* denotes the maximum distance that a vehicle can drive on one full tank of fuel or NGV. In addition, it is calculated by multiplying the fuel consumption by the total amount of fuel in one tank, as shown in Eq. (3):

$$md = fuel_{consumption} \times tank_{capacity}. \tag{3}$$

By contrast, the service cost (*S*) is calculated from taxi rental costs in Thailand, which is approximately 1000 THB/day. We first divided the rental cost by 24 h. After that, we divided the results obtained earlier by 60. Therefore, the service cost per minute is then returned, as in Eq. (4).

$$S = \frac{\frac{taxi_{rental}}{24}}{60}. \tag{4}$$

In addition, *TF* represents the vehicle's fuel consumption when stopped or running slowly for a long time (6 km/h), as shown in Eq. (5).

$$TF = \left(\frac{traffic_{delay} \times 20}{1000}\right) \times fuel_{price}. \tag{5}$$

Note that *PTF* uses the same equation as *TF*. However, *PTF* is the calculation for the traffic delay of the trip before the customer is picked up. However, TF is for the current trip once the customer is onboard. The traffic delay is considered when taxi vehicles drive less than or equal to 6 km/h based on standards by Thailand's land transport department. Then, a factor of 20/1000 is derived from the state in which the vehicle is idling or parking. The vehicle will consume approximately 20 cc of fuel per minute. Therefore, we divide by 1000 to make the unit in liters.

Finally, to calculate the net profit per distance (*C*) of travel, *N* and *d* are used, as in Eq. (6).

$$C = \frac{N}{d}. \tag{6}$$

The net profit per distance obtained using Eq. (6) is summed and divided by the total trip each hour, as shown in Eq. (7). As a result, it returned an average net profit per distance, which this driver earned during each hour of the day. In the final step, it is multiplied by the total distance of travel per hour. Therefore, the net profit per hour (*NPH*) is returned.

$$NPH = \sum_{i=0}^{n} \left(\frac{C_i}{C_n} \right) \times \sum_{i=0}^{m} d_i. \tag{7}$$

Finally, we computed how much fare we need to enable drivers to be more profitable. Unfortunately, drivers are required to spend money daily on the vehicle's energy consumption and car rental. Some costs occur because of the vehicle use.

This cost could generate a profit loss in terms of the actual money obtained (i.e., when a routing occurs). Finally, the top-up amount (*RT*) to be added to the taxi's fare is computed through Eq. (8). This top-up money aims to assist taxi drivers in achieving a more profitable route than in the past. In addition, this outcome can assist the governor in determining reasonable taxi fares.

$$RT = \big((PL \times d) \times Probability\big) + (NT - N), \tag{8}$$

where *RT* represents the top-up money that should be added to the regular fare rate, and *PL* represents the profit loss of each trip, which is calculated by subtracting the net profit of the taxi drivers from the actual profit (result from Eq. (1)). Here, *NT* represents the net profit from a taxi trip with no traffic and the same range of distance.

Finally, the probability that drivers will lose profit from this trajectory pattern (e.g., from the distance of travel or traffic congestion) at a specific time can be computed using the history trajectory data presented in "Data" section.

Profit prediction and result evaluation

In this section, prediction models constructed are described and their accuracy is estimated. This step aims to discover the basic trajectory patterns that are correlated to the changes in the taxi's net profit. The recursive feature elimination (RFE) method [30] is first applied to the data. The RFE is used to select important data features before being input into the predictive model. These features are correlated to the taxi's net profit.

The features used in this study are as follows:

- Features:
    - source direction,
    - destination direction,
    - distance,
    - origin location (latitude and longitude),
    - destination location (latitude and longitude),
    - origin area code (referencing the geospatial 1 km grid),
    - destination area code (referencing the geospatial 1 km grid),
    - trip and waiting times,
    - speed (maximum, minimum, and average),
    - traffic delay,
    - meter status,
    - the time stamp of start and end of the trip,
    - day of the week.

- Predicted variable:
  - Net-profit.

Second, the candidate predictive models include a standard random forest (RF) [31], decision tree (DT) [32], and gradient-boosted regression tree (GBRT). These models were used to reveal the insight provided by the trajectory patterns.

During the model training process, the selected features using RFE are then input into the predictive model. A K-fold cross validation (CV) is used to select the optimal parameters and tune all models. In addition, the CV is set to 5× folds. After obtaining the optimal models, the models are executed. These models aim to predict the net profit for each trajectory from the given features.

Subsequently, the performance of each model is evaluated by comparing the prediction result with actual operational data. This study used real data collected from more than 50 volunteers of taxi drivers for data validation (ground truth). The real data collected consists of the profit (taxi fares from the taxi meter) and expenses for each vehicle determined from the driver's short interview and installation of a developed in-house mobile application, as mentioned in "Taxi survey" section.

Therefore, this information is aggregated into the same format as the features used to train the model earlier. Finally, we verified the result by inputting the ground truth data into the model. The model then predicts the net profit from the given inputs. As a result, if the prediction error is low, it denotes that the model is reliable and statistical for further analysis.

These predictive models were then executed using Apache Spark 2.0. We chose Apache Spark 2.0 to conduct a data analysis for the following reasons:

1. Spark applies analytical tasks in memory. Therefore, it is faster than conducting a data analytic task on MapReduce.
2. Spark supports large-scale and streamed data up to petabytes. Therefore, it is feasible to conduct a real-time analysis from the streamed GPS data points from the taxi vehicle's GPS tracker.

Evaluation of the result

In this study, the root-mean-square error (RMSE), mean absolute error (MAE), accuracy, and f-measure are used to evaluate the model prediction results. In addition, the computational time of these models was also compared. Furthermore, state-of-the-art models from [22, 23] are also reproduced for comparison with our proposed model.

The more accurate prediction results indicate that the selected features are essential and can reveal insights that play a role in the changes in the taxi's profit.

Profit recommendation

In this section, a solution is provided, which consists of the trip type and amount of added fare that is returned from Eq. (8). The recommended action is then presented, and is used to improve the taxi driver's profit and reduce any problems, as described earlier.

In this study, we recommend an add-on fare that enables taxi drivers to receive more profit when driving under different situations. This recommendation is based on the information retrieved from the data, as mentioned in "Data analysis" section. For instance, a trip with traffic and no traffic congestion is a trip with short, medium, and long distances. These different situations are recommended by the specific fare to be added and make the drivers profitable. Therefore, drivers do not need to consider whether the trip they took will reduce their income. The big data model is adapted to changes driven by the data and returns the effective solution at the current time of day.

Experimental results

From the proposed big-data-driven model, the analyzed results are divided into four parts. We first presented a taxi survey. This survey aims to obtain all information related to taxi businesses and how they provide service to customers. These results are preliminary results, and assist in the following part of this study. We then present the experiment conducted on the data and indicate the types of routes that yield more profit. Comparative predictive models are also presented and are used to perform the tasks of extracting insight from the data. Finally, the add-on fare is presented. This is used to recommend a reasonable fare and driving pattern for drivers.

Taxi survey

In this study, a paper-based survey is used to define the taxi problem, making us better understand the taxi problem. As shown in Fig. 7, more than 50 taxi drivers participated in the survey and consisted of various types of vehicles:

- A total of 58% were private taxis, where the driver owns the car.
- A total of 39% were public taxis, where the driver rents the car from a garage.
- The remaining 3% were other types of taxis.

Information was also collected on how taxis find customers and whether they have declined to serve potential passengers.

The results show that taxi drivers look for customers mostly in areas where shopping and department stores are located. This is followed by public transportation such as a train station or bus stop, as shown in Fig. 8.

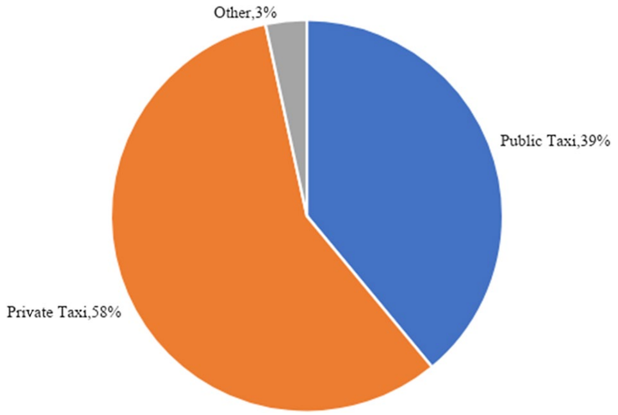

Fig. 7 Taxi types participating in the survey

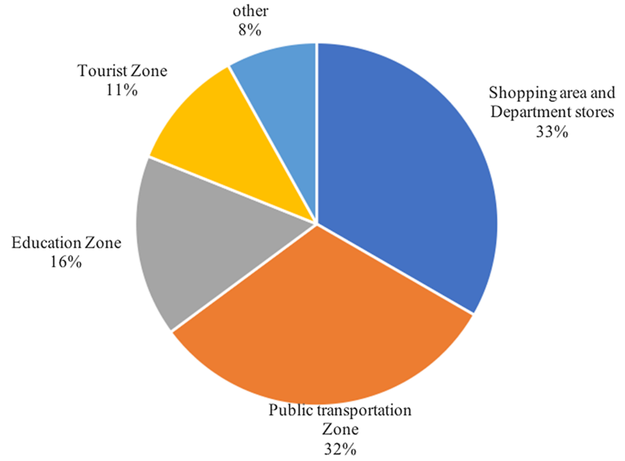

Fig. 8 Location where taxi drivers look for customers in Bangkok

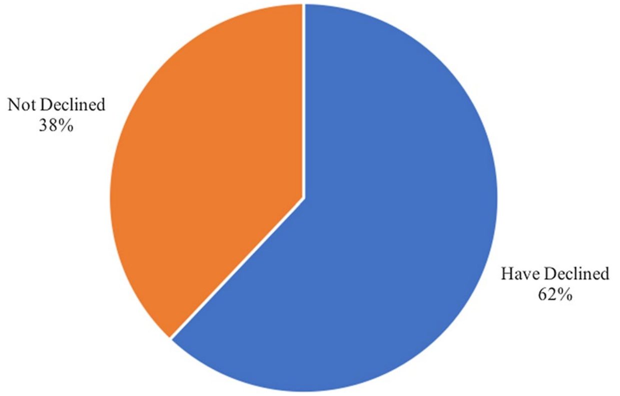

Fig. 9 Proportion of drivers who have declined to serve a customer

Finally, it is essential to discover why taxi drivers decline to serve potential passengers. The survey states that 62% of drivers have experience in declining serve to a customer owing to trip issues (e.g., traffic jams and a time limitation). Consequently, when a customer requests service and the route is different from the desired direction, it may cause a delay in changing shifts with the next driver, as shown in Figs. 9 and 10.

In Fig. 11, an example of a taxi density map from Rama I, Bangkok, which is a popular shopping area at the center of the city, is shown. The figure shows that the high-density areas are within the shopping complex along with Rama I road.

Data analysis

Types of routes that are profitable

The results show that the most profitable journey is when the driver drives in multiple short-distance journeys. This driving style provides more profit than a single average long-distance journey (approximately 49.61%).

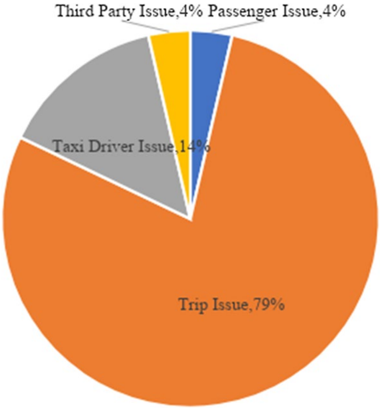

Fig. 10 Reasons why taxi drivers have declined to serve a customer

By contrast, the second-most profitable journey is when drivers drive multiple average medium-distance trips. This driving style enables taxi drivers to earn up to 8.31% more profit than the average long-distance journey.

Finally, when drivers drive multiple average short-distance journeys, it enables drivers to earn up to 38.13% more profit than average medium-distance trips, as shown in Fig. 12.

Important features and predictive model developments

From the results described in the previous section, feature selection based on a recursive feature elimination (RFE) is conducted on the taxi GPS data. This study used an origin–destination (OD) trajectory record extracted from GPS data as the training set. The results of the feature selection by RFE are presented in Fig. 13. After obtaining the necessary features of the predictive models, the top-5 selected features are input into the predictive model.

We aimed to use these classifiers to predict the taxi's profit from the features selected in "Profit prediction and result evaluation" section. Therefore, a lower classifier prediction error indicates that the features selected are appropriate and influence the change in the taxi driver's profit with respect to time.

After the predictive models were executed, the prediction results are as shown in Table 2. The efficiencies of the RF regression, gradient-boosted regression tree (GBRT), and DT were evaluated using the evlautation metrics, as presented in "Evaluation of the result" section. The lower of RMSE, MAE, and the higher of accuracy, and F-measure values indicate that the selected features are efficient in revealing the insight that causes critical changes to the taxi driver's net profit. The model of [22, 23] is also reproduced for comparison with our proposed model.

Note that the profit and net profit shown in this study are denoted in Thai Baht (THB). In addition, for the study by [22], the computational time cannot be compared because it involves an external site survey.

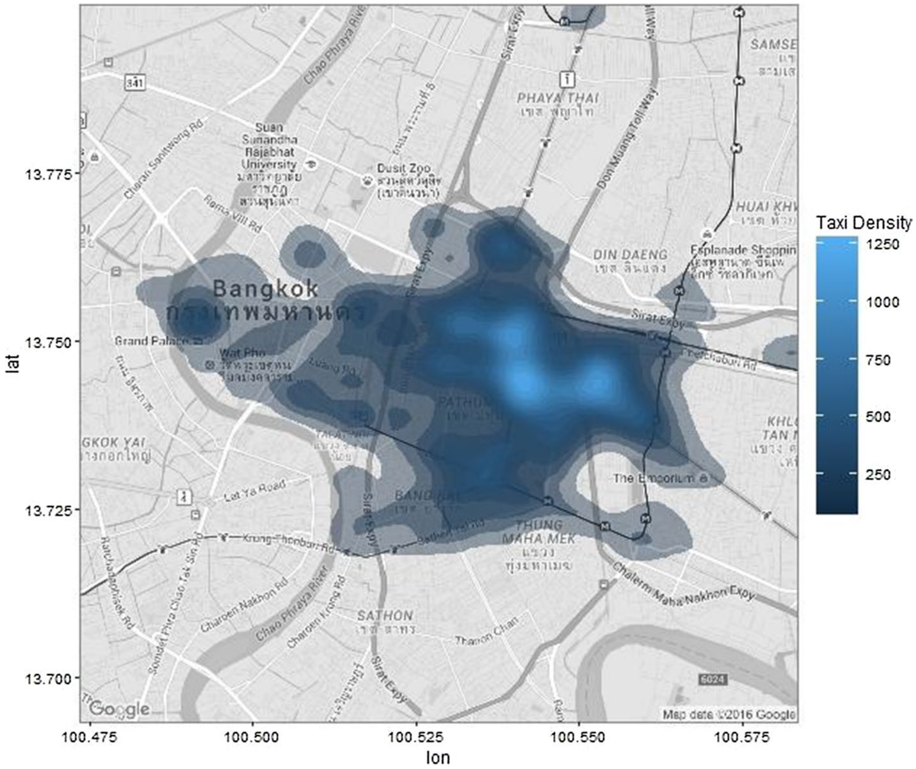

Fig. 11 Taxi density map in Rama I, Bangkok

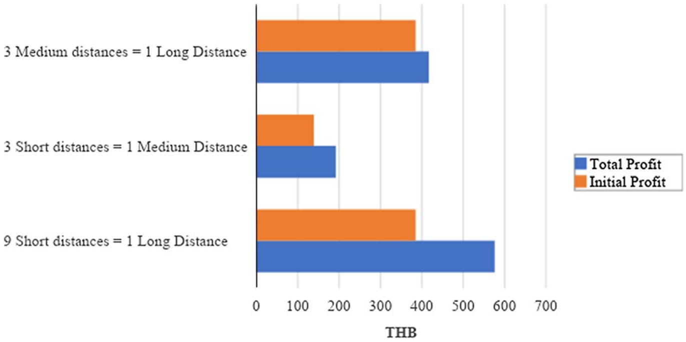

Fig. 12 Comparison of trip distances and profit

Profit recommendation

From the proven features revealed in the previous section, the net profit and profit loss of each trip are calculated. The simulation was conducted for two situations (i.e., typical trips without traffic and trips containing traffic). After this trip information is obtained, the calculation using Eq. 8 is then computed for the amount of fare to be added. Therefore, a reasonable fare is suggested. The solution proposed by taxi drivers is also compared to the solution of our proposed model.

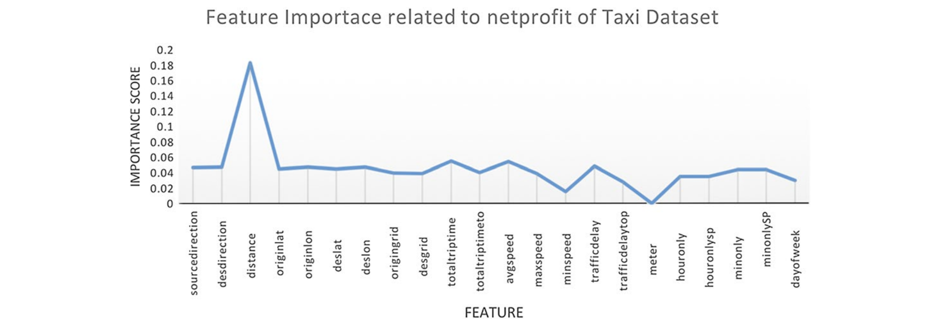

Fig. 13 Feature importance related to net-profit

Table 3 divides the computational results into three categories, i.e., a real trip (no traffic), a real trip (with traffic), and an average trip. Please note that the result in Table 3 is shown in THB.

The distance used in the experiment is set to 5 km, which is the average for a short distance trip. The difference between profit and net profit is also compared. As a result, the amount of profit loss for each trip is returned.

Consequently, we now discover how much the profit that taxi drivers should earn daily differs from the actual amount they earn. These results are revealed from the collected and transformed data, and indicate why the data-driven model is of importance to the governor for making decisions.

After the loss of profit is computed, we then calculate the additional fare needed to help taxi drivers earn more profit using the proposed Eq. (8). This additional fare is presented in the “Additional fare’” rows. In addition, when these additional fares are added to the actual profit, the reasonable fare for the recommendation is return, as presented in the “Recommend fare’” rows.

Table 2 Prediction results

Model	Actual	Predicted	RMSE	MAE	Accuracy (%)	F-meas.	Time (min)
RF	123.27	133.07	9.80	5.51	92.05	0.91	0.19
GBRT	123.27	135.23	11.96	6.01	90.30	0.87	1.33
DT	123.27	135.23	11.96	6.01	90.30	0.87	0.06
[22]	123.27	134.53	11.26	5.65	90.37	0.89	–
[23]	123.27	155.32	32.05	16.11	86.32	0.75	0.67

Table 3 Fare adjustment results

	Real (no traffic) trip	Real (traffic) trip	Average trip
Actual profit	57	57	57
Net profit	46	34	40
Recommend fare	64	72	68
Difference (%)	20	40	30
Additional fare	7	15	7.5

Discussion

After first conducting a site survey, the results are divided into four sections. After obtaining the information, we classified the route followed by determining the important features used to extract insight from the data. Finally, an analysis of fare suggestions is conducted, and the results are discussed here.

Taxi survey

This section describes a paper-based survey conducted to obtain general information from taxi drivers and better understand the problem. The preliminary data of this study are also provided. We discovered that the most popular location to pick up customers is a shopping area or department store, as indicated in Fig. 8. The second most popular area is a transportation area, followed by schools and universities. Further, the survey results show that 62% of the survey respondents have declined to serve customers. The most common reason is an issue with the trip (e.g., a traffic jam or limited time available for driving). This statement is supported by Fig. 10.

If a customer calls for a service, and the route differs from the desired route, it may then cause a delay in changing a shift with the following driver. The drivers are concerned most about the routes and are willing to go to destinations that have yielded a large profit. By contrast, they decline to drive to a destination where they will lose profit or cannot obtain more passengers during the next driving period. The primary assumption is that profits will depend on the taxi trip's routing and distance. Traffic congestion also plays an important role and impacts this problem.

Data analysis

Types of profitable routes

Referring to the previous section, taxi drivers are concerned most about the possibility of obtaining customers and the time limit for each trip. Therefore, in this section, an experiment conducted to reveal profitable trips and prove the earlier assumption is described.

In the experiment, we first extracted and grouped the trip into three groups, namely, "short," "medium," and "long." A trip is classified based on the range of distance. For instance, a short distance is a trip ranging between 1 and 20 km. A medium distance is a trip ranging between 20 and 40 km. More than 40 km is considered a long distance. After every trip is grouped, we computed the average distance, fare, traveling, and times of traffic congestion.

Finally, short, medium, and long trips are compared. For instance, one long-distance trip is 90 km of traveling distance. Therefore, if a short distance is taken for comparison, we take 9 short-distance trips to obtain a total distance equal to that of a long-distance trip.

The experiment shows that frequent average short trips earn more profit than a single average long-distance over the same distance. To increase the taxi driver's performance, we suggest that they consider the destinations that they have serviced.

In general, supposing that the data are unavailable, we cannot know how much profit each trip type provides in a real situation. One possible way is to interview the drivers.

Although this method is acceptable, the information is based on experience and assumptions, and can contain bias.

Unfortunately, some drivers have believed that long-distance trips earn more profit, whereas in reality, more frequent, shorter distance trips increase the net profit. Therefore, this experiment aims to confirm and validate the knowledge of profit earned by drivers before continuing to the later part of the study.

Important features and predictive model developments

In this section, the importance of the features related to profit is determined. The test results show that the most important features evaluated from the RFE are the distance, followed by the travel time, speed, and traffic congestion. After we obtained the essential features, these features were input into the predictive models. The predictive models were then executed.

This study applied only general regression models (e.g., an RF, GBT, and DT) because this study first aims to discover a trajectory pattern that has impacted the taxi driver's profit. Therefore, when selected features are input into the model, the model then uses these features to predict the profit that taxi drivers earned for each trajectory.

Furthermore, we previously chose the trajectory features to represent each movement. These features are the most correlated features of profit. Therefore, if the predictive models are effectively predicted, it is proven that selected features can represent the impact on the taxi profit.

From this outcome, it is unnecessary to perform a more complex model, such as deep learning (DL). The DL model requires more time for hyperparameter tuning and is easy to overfit.

Second, the current ML research has various regression models for applying predictive models. Although the outcomes were acceptable, the RF, GBT, and DT were chosen as the predictive models. These are based on the ensemble method, which combines several models to produce a final optimal predictive model. As a result, it reduces the chance of an overfitting.

Finally, the results of the tree-based predictive models are easy to interpret compared to the neural network or DL. A regression tree aims to partition the input space into several regions by starting with a single root and continuously splitting the node into two child nodes using an appropriate splitting rule. This process continues until the termination conditions are satisfied. Thereafter, each leaf node contains several elements from the training dataset, and a constant prediction value will be assigned to this leaf node by averaging the output values of all of these elements.

The experiment shows that an RF has the best prediction by obtaining a 9.80 RMSE (92.05% of accuracy), followed by [22] with 11.26 RMSE (90.37% of accuracy), the gradient boosted tree and decision tree with an RMSE of 11.96 (90.30% of accuracy). Lastly, [23] with 32.05 RMSE (86.32% of accuracy). The lower RMSE indicates that the insight from the data are accurately extracted and revealed. Consequently, the models accurately returned the predictive results.

From the results, it is shown that the RF provided better results in terms of the prediction. Unfortunately, the computational time is slightly worse than that of the DT. However, the difference was not significant. Therefore, it is still acceptable for practical use.

The state-of-the-art models were also compared. We also reproduced the model from [22, 23] for comparison with our proposed model. The results show that the RF model using this study methodology can outperform the model using the methodology from [22, 23].

The rationale behind this originated from the model designed specifically for the problem of Bangkok city. This includes using the real traffic conditions and internal operational factors of the taxi agency for the computations. These factors are not included in [23]. Nevertheless, this model is typically general and can also be used for any other city problems with minimal changes.

In conclusion, from these predictive model results, it was proven that distance, followed by the travel time, speed, and traffic congestion are crucial factors. They play an essential role in the taxi driver's profit. Therefore, to recommend an add-on fare, it is necessary to consider these factors in the computation. The add-on fare aims to enable driver profit and suggest a reasonable fare to the governor.

Profit recommendation

The previous sections show that distance plays an important role and has a high impact on the taxi driver's net profit. This has led to the recommendation of a fare-rate by using data-driven approaches.

In Bangkok, there is a problem with increasing road traffic congestion as stated in the introduction. All departments try to solve this issue. Furthermore, a group of drivers has introduced some solutions, some of which do not meet customer satisfaction. For this reason, a solution that can help taxi drivers earn more profit than in the past is proposed, and is compared with the solution proposed by taxi drivers.

The analysis results show that a shorter distance taxi ride will earn more profit. As a result, the results suggest that drivers should take more short-distance trips. This is because they will have more benefits than when waiting for a long amount of time for a long-distance trip and without the risk of not finding the next passenger. Next, the taxi fare is calculated when driving under a situation with no traffic and when traffic is introduced. Fortunately, a cost–distance algorithm can be used to calculate this process.

The trajectory features obtained are input to calculate the fare, the difference in fare, and the profit loss (i.e., caused by traffic congestion, distance, and vehicle costs). Subsequently, the occurrence of the selected driving pattern is computed. Finally, a value-added fare is suggested using Eq. (8).

The results demonstrate that for an actual short-distance trip of 5 km with a traffic delay of 10 min, for example, the customer pays 77 THB. Furthermore, the result also demonstrates that traffic congestion reduces the driver's profit from 57 to 34 THB (an approximately 40% decrease). This result proved that not only does the distance of the trip play a role in changing the driver's profit, traffic congestion is also a crucial factor involving such changes.

However, in a new solution proposed by taxi drivers, the customer pays 127 THB, which is a 65% increase. By contrast, according to our recommendation using Eq. (8), customers pay only 72 THB, 55 THB less than the solution proposed by drivers. In other words, this is an optimal fare that provides benefits to both drivers and customers.

However, it does not significantly change between the net profit and profit in a short-distance trip without traffic. It is therefore unnecessary to make changes to the fare.

In conclusion, the recommendation regarding the taxis fare is driven by large-scale taxi trajectory data. Furthermore, this study also shows that unmeaningful data can be transformed into valuable information and knowledge that significantly improve the taxi services in urban cities. If there is a lack of data, it is difficult to recommend reasonable fares for taxis drivers and support the decision-making of the governors.

Conclusions

This study investigated several taxi problems and improved the methodology of [23]. The primary purpose is to determine why a taxi driver declines to serve customers in Bangkok using a data-driven approach. Taxi probe data collected over several months in the Bangkok area were used for analysis.

To begin with, we started with data exploration on raw data, and cleaned unwanted data included outliers. The technique behind this process is geospatial techniques such as geospatial grids and intersections. We then constructed a big-data-driven model to be a framework for the analysis of this data.

From the results, there are two main findings. First, it was discovered that the solution proposed by taxi drivers directly benefits the drivers but does not satisfy customers. This study suggests that the recommendation returned from the proposed model will satisfy both customers and drivers. The number of fares to be added is a balance between the traditional fare and the fee proposed by the driver's so-called optimal solution.

The driver can make more profit, and customers do not have to pay an extremely high fare under traffic congestion.

Second, it was proven that distance, followed by travel time, speed, and traffic congestion, are crucial factors for determining the trajectory patterns. These patterns have a significant impact on taxi driver profits. As a result, these factors are considered for recommending the add-on fare to the drivers. The recommendation of an add-on fare is based on travel distance and traffic congestion at a specific time. These factors are extracted from big data and are effectively predicted by the RF, GBRT, and DT at up to 9.80 RMSE and 0.19 min of computational time.

In this study, we demonstrated the practical use of the proposed model through an example of a regular taxi trip. The results show the solution when the trip is involved and not involved under traffic congestion. As a result, the recommendation of an add-on fare can be used to enable taxi drivers to earn more profit. For instance, 7 THB is added for a non-traffic congestion trip and 15 THB is added for the traffic congestion trip. Therefore, drivers do not need to consider whether the trip they received will reduce their income. The big data model is adapted to changes driven by the data and returns the effective solution at the current time of the day.

Finally, this study has certain limitations. It only considered the routing perspective. It is assumed that the operational cost is equal among vehicles. However, other perspectives also need to be considered (e.g., economic, human resources, and taxi agency operation).

In a future study, we will improve the proposed model to handle more complex scenarios and correlated factors. A comparison of the deep learning approach is also considered to support the dynamic system when suggesting routing and optimizing profits.

Abbreviations

CV: Cross-validation; DT: Decision trees; DL: Deep learning; GOTR: Global-optimal trajectory retrieving; GBRT: Gradient-boosted regression tree; h: Hour; IMEI: International Mobile Equipment Identity; km: Kilometer; LOS: Level of service; ML: Machine learning; MAE: Mean-absolute error; NGV: Natural gas for vehicle; OD: Origin and destination; RF: Random forest; RFE: Recursive feature elimination; RMSE: Root-mean-square error.

Acknowledgements

This research was supported in part by the Scholarship for Excellent Thai Students from Sirindhorn International Institute of Technology, Thammasat University, no. ETS-G-S1Y17/061. The GPS data used in this study were provided by Toyota Tsusho Electronic (Thailand) Co., Ltd. We would like to thank our colleagues from Bhumipalang Taxi Co., Ltd., who provided insight and expertise that greatly assisted our study.

Authors' contributions

TP and TH conceived the idea. TP developed the theory and performed computations. TH encouraged TP to investigate mobility and economic aspects and supervised the findings of this study. All authors discussed the results and contributed to the final manuscript. All authors read and approved the final manuscript.

Funding

The authors received no specific funding for this work.

Availability of data and materials

The data that support the findings of this study are available from Toyota Tsusho Electronic (Thailand) Co., Ltd., but restrictions apply to their availability, which were used under license for the current study, and thus are not publicly available. However, the data are available from the authors upon reasonable request and with permission from Toyota Tsusho Electronic (Thailand) Co., Ltd.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand. [2] School of Knowledge Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan.

Received: 21 September 2020 Accepted: 9 December 2020
Published online: 06 January 2021

References

1. Jayasooriya SACS, Bandara YMMS. Measuring the economic costs of traffic congestion. In: 3rd international Moratuwa engineering research conference, MERCon 2017; 2017. p. 141–6. https://doi.org/10.1109/MERCon.2017.7980471.
2. Wade R. World taxi prices: what a 3-kilometer ride costs in 88 big cities; 2017. https://www.priceoftravel.com/555/world-taxi-prices-what-a-3-kilometer-ride-costs-in-72-big-cities/.
3. Numbeo: cost of living in Bangkok. https://www.numbeo.com/cost-of-living/in/Bang-Kruai-Thailand.
4. Zheng Y, Zhang L, Xie X, Ma W-Y. Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th international conference on World Wide Web—WWW '09; 2009. p. 791. https://doi.org/10.1145/1526709.1526816. http://portal.acm.org/citation.cfm?doid=1526709.1526816.
5. Yue Y, Zhuang Y, Li Q, Mao Q. Mining time-dependent attractive areas and movement patterns from taxi trajectory data. Knowledge creation diffusion utilization. Thousand Oaks: Sage Publications; 2009.
6. Hwang R-H, Hsueh Y-L, Chen Y-T. An effective taxi recommender system based on a spatio–temporal factor analysis model. Inf Sci. 2015;314(4):28–40. https://doi.org/10.1016/j.ins.2015.03.068.
7. Qu M, Zhu H, Liu J, Liu G, Xiong H. A cost-effective recommender system for taxi drivers. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining—KDD '14; 2014. p. 45–54. https://doi.org/10.1145/2623330.2623668.
8. Kamimura J, Ogawa M, Wakayama H, Iga N, Shiota N, Yano M. D-Taxi: adaptive area recommendation system for taxis by using DiRAC. In: Proceedings of 2013 international conference on connected vehicles and expo, ICCVE 2013, vol. 4; 2013. p. 507–8. https://doi.org/10.1109/ICCVE.2013.6799845.
9. Ding Y, Liu S, Pu J, Ni LM. HUNTS: a trajectory recommendation system for effective and efficient hunting of taxi passengers. In: Proceedings of IEEE international conference on mobile data management, vol. 1; 2013. p. 107–16. https://doi.org/10.1109/MDM.2013.21.
10. Zhang M, Liu J, Liu Y, Hu Z, Yi L. Recommending pick-up points for taxi-drivers based on spatio–temporal clustering. In: Proceedings of the 2nd international conference on cloud and green computing and 2nd international conference on social computing and its applications, CGC/SCA 2012; 2012. p. 67–72. https://doi.org/10.1109/CGC.2012.34.
11. Yuan J, Zheng Y, Zhang L, Xie X, Sun G. Where to find my next passenger. In: Proceedings of the 13th international conference on ubiquitous computing; 2011. p. 109–18. https://doi.org/10.1145/2030112.2030128.

12. Zhang D, He T. PCruise: reducing cruising miles for taxicab networks. In: Proceedings of the real-time systems symposium; 2012. p. 85–94. https://doi.org/10.1109/RTSS.2012.61.
13. Qi G, Pan G, Li S, Wu Z, Zhang D, Sun L, Yang LT. How long a passenger waits for a vacant taxi? Large-scale taxi trace mining for smart cities. In: Proceedings of the 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, vol. 4; 2013. p. 1029–36. https://doi.org/10.1109/GreenCom-iThings-CPSCom.2013.175.
14. Zhang Y, Haghani A. A gradient boosting method to improve travel time prediction. Transp Res C Emerg Technol. 2015;58:308–24. https://doi.org/10.1016/j.trc.2015.02.019.
15. Moreira-Matias L, Gama J, Ferreira M, Mendes-Moreira J, Damas L. Predicting taxi—passenger demand using streaming data. IEEE Trans Intell Transp Syst. 2013;14(3):1393–402. https://doi.org/10.1109/TITS.2013.2262376.
16. Yingjun Y, Cui H, Shaoyang Z, Yingjun Y. A prediction model of the number of taxicabs based on wavelet neural network. Procedia Environ Sci. 2012;12:1010–6. https://doi.org/10.1016/j.proenv.2012.01.380.
17. Zhou C, Dai P, Wang F, Zhang Z. Predicting the passenger demand on bus services for mobile users. Pervasive Mob Comput. 2016;25(2013):48–66. https://doi.org/10.1016/j.pmcj.2015.10.003.
18. Phiboonbanakit T, Horanont T. How does taxi driver behavior impact their profit? Discerning the real driving from large scale GPS traces. In: UbiComp 2016 adjunct—proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing. 2016. https://doi.org/10.1145/2968219.2968417.
19. Bai YW, Wang EW. Design of taxi routing and fare estimation program with re-prediction methods for a smart phone. In: 2012 IEEE I2MTC—proceedings of the international instrumentation and measurement technology conference; 2012. p. 716–21. https://doi.org/10.1109/I2MTC.2012.6229165.
20. Egan M, Jakob M. Market mechanism design for profitable on-demand transport services. Transp Res B Methodol. 2016;89:178–95. https://doi.org/10.1016/j.trb.2016.04.020.
21. Zhang W, Honnappa H, Ukkusuri SV. Modeling urban taxi services with e-hailings: a queueing network approach. Transp Res Procedia. 2018;38:751–71. https://doi.org/10.1016/j.trpro.2019.05.039.
22. Wong RCP, Szeto WY. An alternative methodology for evaluating the service quality of urban taxis. Transp Policy. 2018;69:132–40. https://doi.org/10.1016/j.tranpol.2018.05.016.
23. Čulík K, Kalašová A, Otahálová Z. Alternative taxi services and their cost analysis. Transp Res Procedia. 2020;44:240–7. https://doi.org/10.1016/j.trpro.2020.02.047.
24. THAIest: Bangkok taxi—fare & other Thailand taxi meter tips; 2020. https://thaiest.com/thailand/bangkok/taxi.
25. Geofabrik: OpenStreetMap data extracts. https://download.geofabrik.de.
26. Postgis.net: PostGIS 2.3.11dev manual. Technical report, Postgis.net; 2020.
27. Phiboonbanakit T, Horanont T. Who will get benefit from the new taxi fare rate? Discerning the real driving from Taxi GPS data. In: 7th international conference on information communication technology for embedded systems 2016, IC-ICTES 2016; 2016. https://doi.org/10.1109/ICTEmSys.2016.7467125.
28. Phiboonbanakit T. Analyzing Bangkok city taxi ride: ReformingFares for profit sustainability using Big DataDriven model. 2020. https://github.com/nuttthp/Analyzing-Bangkok-City-Taxi-Ride-ReformingFares-for-Profit-Sustainability-using-Big-DataDriven-Mode.git.
29. Phiboonbanakit T. Research project on problems and the impact of fares on current taxi drivers' occupation. 2020. https://forms.gle/sVQ4YJCKw8TAERSJ6.
30. Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemom Intell Lab Syst. 2006;83(2):83–90. https://doi.org/10.1016/j.chemolab.2006.01.007.
31. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(December):18–22. https://doi.org/10.1177/154405910408300516.
32. Lemm S, Blankertz B, Dickhaus T, Müller KR. Introduction to machine learning for brain imaging. NeuroImage. 2011;56(2):387–99. https://doi.org/10.1016/j.neuroimage.2010.11.004.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.