

การพยากรณ์ระยะเวลาการเดินทางด้วยวิธีการเรียนรู้ของเครื่อง โดยใช้ข้อมูลจากพาหนะ และโทรศัพท์มือถือในจังหวัดกรุงเทพมหานคร

ปวริศ เวชวรรณกิจกุล*¹ และ พิศิษฐ์ จารุมณีโรจน์²

ภาควิชาวิศวกรรมอุตสาหการ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพฯ 10330

Received: 2 May 2022; Revised: 4 October 2022; Accepted: 25 October 2022

บทคัดย่อ

การขนส่งไปยังลูกค้าปลายทาง (Last mile delivery) ถือเป็นหนึ่งในกระบวนการขนส่งสินค้าที่สำคัญ ที่ครอบคลุมต้นทุนกว่าร้อยละ 75 ของห่วงโซ่อุปทาน อย่างไรก็ตาม การขนส่งในกระบวนการดังกล่าวกลับมีประสิทธิภาพที่ค่อนข้างต่ำ อันเป็นผลเนื่องจากความไม่แน่นอนของระยะเวลาการเดินทางที่แปรเปลี่ยนไประหว่างวัน เพื่อเพิ่มประสิทธิภาพของกระบวนการดังกล่าว ผู้วิจัยจึงได้ทำการพัฒนาต้นแบบการพยากรณ์ระยะเวลาการเดินทางระหว่างจุดด้วยวิธีการเรียนรู้ของเครื่อง โดยอาศัยข้อมูลตำแหน่งพาหนะ และโทรศัพท์มือถือของรถแท็กซี่ ที่เก็บโดยมูลนิธิศูนย์ข้อมูลจราจรอัจฉริยะไทย (iTic) เป็นกรณีศึกษา ทั้งนี้ ผู้วิจัยได้ทำการจำกัดขอบเขตของการศึกษา โดยคัดเลือกเฉพาะข้อมูลของรถแท็กซี่ที่กำลังรับส่งผู้โดยสาร ที่มีพิกัดอยู่ภายในบริเวณกรุงเทพมหานครจำนวน (4.8 ล้าน) ข้อมูล ไปสร้างต้นแบบการพยากรณ์ด้วยวิธีการเรียนรู้ของเครื่องผ่านอัลกอริทึมแบบต่าง ๆ โดยใช้วิธี k-fold cross validation ในการแบ่งข้อมูลเพื่อลดปัญหา overfitting ผู้วิจัยพบว่าต้นแบบที่มีประสิทธิภาพสูงสุด คือ ต้นแบบที่พัฒนาขึ้นจากอัลกอริทึม XGBoost ซึ่งมีค่า RMSE อยู่ที่ 163.3069 และใช้เวลาในการสร้างต้นแบบเพียง 14.51 วินาที อย่างไรก็ตาม ต้นแบบที่พัฒนาขึ้นด้วยอัลกอริทึม LightGBM และ CatBoost ก็มีแนวโน้มที่ดีในการนำไปพัฒนาต่อ เนื่องจากให้ค่า RMSE และใช้เวลาในการสร้างต้นแบบที่ค่อนข้างค่อนข้างต่ำ

คำสำคัญ: การพยากรณ์, ระยะเวลาการเดินทาง, วิธีการเรียนรู้ของเครื่อง, ข้อมูลจากพาหนะ และโทรศัพท์มือถือ

* Corresponding author. E-mail: pawaris.wach@gmail.com

¹ ภาควิชาวิศวกรรมอุตสาหการ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

² ภาควิชาวิศวกรรมอุตสาหการ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

A Prediction of Travel Time by Machine Learning Approaches with Mobile Probe Data in Bangkok

Pawaris Wachwanakijkul^{*1} and Pisit Jarumaneeroj²

Department of Industrial Engineering, Faculty of Engineering, Chulalongkorn University,
Bangkok 10330

Received: 2 May 2022; Revised: 4 October 2022; Accepted: 25 October 2022

Abstract

Last mile delivery is one of main transportation processes that accounts for more than 75% of total supply chain cost; however, the efficiency of such a process is relatively low, due largely to uncertainty in travel times across the day. In order to improve this process, various machine-learning based approaches are herein investigated for the development of more accurate travel time prediction models, using mobile probe data, collected by ITic, as a case study. For ease of implementation, we have preselected the information of taxis that currently provide services within the area of Bangkok (about 4.8 million records) for the construction of travel time prediction models. K-fold cross validation is also adopted to help reduce overfitting issues. Our results indicate that XGBoost is the most effective algorithm that provides the least RMSE (166.3069), while spending only 14.51 seconds in the model construction phase. Nonetheless, LightGBM and CatBoost seem to have good potentials for further studies as they provide relatively low RMSEs and computational times, when compared to other machine-learning approaches.

Keywords: prediction, travel time, machine learning, mobile probe data

* Corresponding author. E-mail: pawaris.wach@gmail.com

1 Department of Industrial Engineering, Faculty of engineering, Chulalongkorn University

2 Department of Industrial Engineering, Faculty of engineering, Chulalongkorn University

1. บทนำ

ประเทศไทยมีผู้ให้บริการธุรกิจขนส่งทางถนนมากมาย ไม่ว่าจะเป็นสินค้าประเภท คน พัสดุ หรือ อาหาร และ ในการขนส่งสินค้าเหล่านี้ จะมีกระบวนการหนึ่งที่เหมือนกัน คือ การขนส่งไปยังลูกค้าปลายทาง (Last mile delivery) ซึ่งเป็นการรับสินค้าจากสถานที่หนึ่งไปส่งยังอีกสถานที่หนึ่ง กระบวนการดังกล่าว เป็นหนึ่งในส่วนสำคัญของธุรกิจขนส่งทางถนน ซึ่งโดยทั่วไปแล้วมักมีประสิทธิภาพต่ำ และมีต้นทุนสูง [1] ทั้งนี้ในส่วนของการขนส่งไปยังลูกค้าปลายทางมีค่าระหว่างร้อยละ 13 ถึงร้อยละ 75 ของโซ่อุปทาน (Supply Chain) ทั้งหมด [2] ส่วนประสิทธิภาพของการขนส่งไปยังลูกค้าปลายทางนั้นขึ้นอยู่กับหลายปัจจัย เช่น ความหนาแน่นของผู้บริโภคและกรอบเวลา [3] ความชำนาญเส้นทางของพนักงานขนส่ง [4] และ การนำจ่ายพัสดุไม่สำเร็จ [2] เป็นต้น

มูลค่าตลาดขนส่งพัสดุในประเทศไทย ตั้งแต่ปี พ.ศ. 2558 ถึง พ.ศ.2563 เติบโตขึ้นทุกปี [5] การวางแผนเส้นทางเดินรถเพื่อขนส่งสินค้าจึงมีส่วนสำคัญในการเพิ่มประสิทธิภาพของการขนส่งไปยังลูกค้าปลายทาง แรกเริ่มในงานวิจัยส่วนใหญ่มักมุ่งเน้นไปที่การแก้ปัญหาวิธีสั้นสุด (Shortest path) จากข้อมูลชุดเดิมเพียงชุดเดียว [6] ต่อมาจึงเริ่มมีการพัฒนาปัญหาการจัดเส้นทางในรูปแบบเดิม ให้มีความเสมือนจริงมากขึ้นเพื่อรับมือกับเงื่อนไขสภาพแวดล้อมที่เปลี่ยนแปลงไปอยู่ตลอดเวลา เช่น ความหนาแน่นของการจราจรในแต่ละช่วงเวลาและสถานที่ ที่ส่งผลกระทบต่อระยะเวลาในการเดินทาง [7]

กรุงเทพมหานคร เป็นเมืองหลวงของประเทศไทย ซึ่งมีอันดับระดับความแออัดของสภาพการจราจรอยู่ในอันดับที่ 10 จาก 416 เมือง 57 ประเทศ จากการจัดอันดับดังกล่าว สามารถกล่าวได้ว่า กรุงเทพมหานครนั้นเป็นเมืองที่มีสภาพการจราจรแออัดมาก [8] ความแปรปรวนของความแออัดของสภาพการจราจรในกรุงเทพมหานครมีความแตกต่างกันในแต่ละช่วงเวลาของแต่ละวันในสัปดาห์ ส่งผลทำให้ระยะเวลาในการเดินทางจากสถานที่หนึ่งไปยังอีกสถานที่หนึ่ง อาจแตกต่างกันมากในแต่ละช่วงเวลา

จากการค้นคว้างานวิจัยการจัดเส้นทางที่นำความไม่แน่นอนของเวลาเดินทางเข้ามาประกอบด้วย (Online or

real-time problem) พบว่า มีหลายเทคนิคที่ถูกนำมาใช้ในส่วนของการหาระยะเวลาเดินทางที่เปลี่ยนแปลงไปในแต่ละช่วงเวลา เช่น การใช้แบบจำลองแบบสุ่ม (Stochastic Model) [9] การเรียกข้อมูลจากแอปพลิเคชัน Google Map (Google distance Matrix API) [10] การวิเคราะห์ข้อมูลที่รวบรวมได้จากสื่อโซเชียล [11] อย่างไรก็ตาม งานวิจัยเหล่านี้ไม่ได้กล่าวถึงรายละเอียดของข้อมูลที่น่าสนใจ หรือความแม่นยำของระยะเวลาเดินทางมากนัก ทำให้ไม่สามารถนำมาใช้เปรียบเทียบผลคำตอบกับงานวิจัยประเภทเดียวกันได้ นอกจากนี้แม้ว่าจะสามารถใช้การเรียกข้อมูลเมตริกระยะเวลาเดินทาง จากแอปพลิเคชันผู้ให้บริการแผนที่ต่าง ๆ เช่น Google Map, TomTom, Apple Map, Longdo Traffic, Nostra Map การเรียกข้อมูลในปริมาณที่มากกว่าข้อจำกัด จำเป็นต้องเสียค่าใช้จ่าย หรือหากเรียกข้อมูลจากแอปพลิเคชัน MapQuest ซึ่งเป็นผู้ให้บริการแผนที่แบบไม่มีค่าใช้จ่าย และไม่มีข้อจำกัดในด้านปริมาณการเรียกข้อมูล MapQuest กลับมีข้อจำกัดในด้านขนาดของเมตริก ทำให้ไม่สามารถเรียกข้อมูลครั้งละมาก ๆ ได้

ผู้วิจัยจึงมีความสนใจในการสร้างต้นแบบการคาดการณ์ระยะเวลาเดินทางระหว่างพิกัดสองจุดด้วยวิธีการเรียนรู้ของเครื่อง (Machine learning) ในเขตพื้นที่ของจังหวัดกรุงเทพมหานครจากชุดข้อมูลที่สามารถเข้าถึงได้โดยสาธารณะ เพื่อลดข้อจำกัดดังกล่าว

2. การทบทวนวรรณกรรม

การคาดการณ์เวลาเดินทาง เป็นหนึ่งในส่วนสำคัญของระบบ ITS (Intelligent transportation system) ในช่วง 30 ปีที่ผ่านมา นักวิจัยด้านการขนส่ง และ นักวิทยาศาสตร์ข้อมูลได้พัฒนาหลากหลายเทคนิคเพื่อเพิ่มความน่าเชื่อถือแก่วิธีการคาดการณ์เวลาเดินทางในอนาคต [18] โดยทั่วไปแล้วเทคนิคเหล่านั้นสามารถแบ่งออกได้เป็น 3 กลุ่ม ได้แก่ Naive methods, traffic theory-based method และ data-driven methods โดยเทคนิค data-driven นั้นจะใช้ชุดข้อมูลที่เกี่ยวข้องกับการเดินทาง และ ข้อมูลวิธีทางสถิติในการคาดการณ์เวลาเดินทาง ทำให้มีผู้ที่สามารถวิจัยหัวข้อนี้ได้เยอะขึ้น เพราะไม่จำเป็นต้องมีความเชี่ยวชาญในด้านทฤษฎีการจราจรโดยเฉพาะ [19] เทคนิคนี้

จำเป็นต้องใช้ข้อมูลปริมาณมากซึ่งอาจจะไม่ได้มีอยู่ในทุกที่ ต้นแบบที่สร้างจากเทคนิคนี้จะมีความแม่นยำมากขึ้นจาก ความพร้อมของข้อมูล [20]

ถึงแม้ว่าจะมีงานวิจัยมากมายที่เกี่ยวข้องกับการ คาดการณ์ระยะเวลาเดินทางบนท้องถนน แต่มีไม่มากนักที่ นำวิธีการเรียนรู้ของเครื่องมาใช้ร่วมกัน ซึ่งมีข้อดีในการหา รูปแบบที่ซับซ้อนของข้อมูลด้วยอัลกอริทึมที่มีความซับซ้อน มากกว่าวิธีแบบดั้งเดิมทำให้เกิดการใช้ทรัพยากร คอมพิวเตอร์อย่างมีประสิทธิภาพ และยังสามารถใช้กับชุด ข้อมูลขนาดใหญ่ได้อีกด้วย และงานวิจัยประเภทนี้ยังมีการ ใช้ชุดข้อมูลจากแหล่งข้อมูลที่แตกต่างกัน เช่น ข้อมูล สาธารณะของแท็กซี่ในนิวยอร์ก (NYC Yellow Cab trip record) [12] เว็บไซต์วางแผนการเดินทาง Uber Movement [13] ซึ่งเป็นข้อมูลที่บันทึกจากแต่ละประเทศ ซึ่งมีความแตกต่างในการวิเคราะห์โครงข่ายถนน พฤติกรรม การขับขี่ ความหนาแน่นของการจราจร พื้นที่ภูมิประเทศ ส่งผลทำให้การคาดการณ์ระยะเวลาเดินทางบนท้องถนน หนึ่ง ๆ ไม่สามารถนำไปประยุกต์ใช้กับการคาดการณ์ ระยะเวลาอื่น ๆ ได้อย่างมีความแม่นยำ งานวิจัยประเภท เดียวกันจึงมีความแตกต่างกันอย่างเห็นได้ชัดในด้านของชุด ข้อมูล และ วิธีการที่ใช้คาดการณ์ ซึ่งงานวิจัยที่ใช้วิธีการ เรียนรู้ของเครื่องส่วนใหญ่จะหยิบยกอัลกอริทึมบางอย่าง สำหรับวิเคราะห์การถดถอยมาใช้ เช่น Linear Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF) [12-14] โดยการคัดเลือกอัลกอริทึมเหล่านี้มาใช้คาดการณ์ก็ เป็นอีกหนึ่งความแตกต่างของงานวิจัยในกลุ่มประเภทนี้

3. ระเบียบวิธี

3.1 วิธีการเรียนรู้ของเครื่อง

วิธีการเรียนรู้ของเครื่อง มีความหมายโดยกว้างว่า เป็นวิธีการคำนวณด้วยคอมพิวเตอร์โดยใช้ประสบการณ์ ในการเพิ่มประสิทธิภาพ หรือทำให้การคาดการณ์แม่นยำขึ้น โดยประสบการณ์ในที่นี้หมายถึง ข้อมูลสารสนเทศที่ผ่านมา ในอดีต ซึ่งโดยปกติจะถูกจัดเก็บอยู่ในรูปแบบข้อมูล อิเล็กทรอนิกส์ ข้อมูลเหล่านี้จะถูกดัดแปลงให้อยู่ในรูปแบบ ของตัวเลขเชิงปริมาณ แล้วทำการจัดแบ่งประเภทของข้อมูล โดยมนุษย์เพื่อให้สามารถนำไปใช้ในวิธีการเรียนรู้ของเครื่อง

ได้ ทั้งนี้คุณภาพ และขนาดของข้อมูลมีความสำคัญต่อ ประสิทธิภาพของการคาดการณ์ที่ถูกสร้างโดยวิธีการเรียนรู้ ของเครื่อง [15]

วิธีการเรียนรู้ของเครื่องนั้นถูกออกแบบมาเพื่อใช้ กับงานหลากหลายรูปแบบ ในงานวิจัยนี้จะใช้วิธีการเรียนรู้ ของเครื่องกับปัญหาการถดถอย (Regression) ซึ่งเป็นปัญหา การคาดการณ์ ค่าที่แท้จริงของสิ่งของแต่ละอย่าง ตัวอย่างเช่น การคาดการณ์มูลค่าของหุ้นหรือความผันแปร ของตัวแปรทางเศรษฐกิจ โดยความผิดพลาดของการถดถอย นี้จะขึ้นอยู่กับขนาดของความแตกต่างระหว่างค่าที่เกิดขึ้น จริงและค่าที่เกิดจากการคาดการณ์

ประเภทของวิธีการเรียนรู้ของเครื่องจะถูกแบ่งโดย ข้อมูลที่ใช้ ซึ่งในงานวิจัยนี้จะเป็นการเรียนรู้แบบมีผู้สอน (Supervised learning) กล่าวคือ ผู้เรียนรู้จะได้รับข้อมูลที่มี คำตอบอยู่แล้ว และทำการคาดการณ์เพื่อเปรียบเทียบกับ คำตอบนั้น

3.2 การคัดเลือกและการประเมินผลของต้นแบบ

ในการสร้างต้นแบบการคาดการณ์ระยะเวลา เดินทางบนท้องถนนระหว่างพิกัดสองจุดด้วยวิธีการเรียนรู้ ของเครื่อง จำเป็นต้องมีการคัดเลือกต้นแบบที่ดีที่สุดไม่ว่าจะ เป็นในการสร้างต้นแบบเดียวกัน หรือระหว่างต้นแบบอื่น ๆ ซึ่งในงานวิจัยนี้จะใช้เมตริก RMSE (Root Mean Square Error) เป็นตัวชี้วัดประสิทธิภาพของต้นแบบ โดยสามารถ คำนวณได้จากค่าความผิดพลาดระหว่างค่าการคาดการณ์ และ ค่าของข้อมูลที่ใส่เข้าไปให้กับผู้เรียนรู้ซึ่งจะมีค่าตั้งแต่ 0 ถึง ∞ ดังแสดงในสมการที่ 1 หากค่า RMSE มีค่าต่ำจะ หมายความว่า ต้นแบบนั้น ๆ มีประสิทธิภาพที่สูง

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

โดยที่ $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ คือ ค่าที่คาดการณ์ได้จากต้นแบบ

$y_1, y_2, \dots, y_n$ คือ ค่าที่ใส่ให้กับผู้เรียนรู้

n คือ จำนวนของค่าที่ใส่เข้าไปให้กับผู้เรียนรู้

3.3 การลดความผิดพลาดที่เกิดขึ้นจากการสร้างต้นแบบ

ปัญหาที่เกิดขึ้นในการสร้างต้นแบบเป็นปัญหาที่ เกิดจากโครงสร้างของวิธีการเรียนรู้ของเครื่อง ซึ่งหากสร้าง

ต้นแบบจากข้อมูลชุดหนึ่งโดยตั้งจุดประสงค์ในการทำให้ค่าผิดพลาดมีค่าน้อยมาก ๆ เมื่อนำต้นแบบนี้ไปใช้กับข้อมูลชุดอีกชุดหนึ่งซึ่งผู้เรียนรู้ไม่เคยเห็นมาก่อนอาจมีความเสี่ยงที่ค่าผิดพลาดจะมากขึ้น เพราะตัวต้นแบบเกิดการจำลักษณะเฉพาะของข้อมูลชุดแรกมากเกินไป ซึ่ง โดยทั่วไปเรียกปัญหานี้ว่า overfitting [16]

งานวิจัยนี้ได้แก้ไขปัญหการเกิด overfitting กับต้นแบบโดยใช้วิธี K-fold cross validation หลักการของวิธีนี้ คือ การแบ่งข้อมูลเป็น K ส่วนแบบสุ่ม เพื่อให้ข้อมูลมีการกระจายตัว แล้วทำการวัดประสิทธิภาพของต้นแบบโดยใช้ข้อมูล 1 ส่วน (test set) สำหรับส่วนที่เหลือจะนำมาใช้ในการสร้างต้นแบบ (train set) ดำเนินการซ้ำ โดยสลับ test set จนครบ K ส่วน และหาต้นแบบที่มีประสิทธิภาพมากที่สุด วิธีนี้จะทำให้เกิดการสร้าง และทดสอบต้นแบบหลายครั้ง ส่งผลทำให้ปัญหการเกิด overfitting ลดลง [17]

3.4 PyCaret

PyCaret คือ โลเบรารีที่รวบรวมโลเบรารีอัลกอริทึมต่าง ๆ ของวิธีการเรียนรู้ของเครื่อง ใช้ในการเปรียบเทียบต้นแบบที่เกิดจากอัลกอริทึมเหล่านั้น โดยโลเบรารี PyCaret สามารถเปรียบเทียบอัลกอริทึมการเรียนรู้ของเครื่องแบบถดถอยได้ถึง 20 อัลกอริทึมดังนี้ Bayesian Ridge, Lasso Regression, Orthogonal Matching Pursuit, Decision Tree Regressor, AdaBoost Regressor, Elastic Net, K Neighbors Regressor, Lasso Least Angle Regression, Dummy Regressor, Passive Aggressive Regressor, Huber Regressor, Extreme Gradient Boosting, Random Forest Regressor, Extra Trees Regressor, Light Gradient Boosting Machine, CatBoost Regressor, Gradient Boosting Regressor, Linear Regression, Ridge Regression, Least Angle Regression

3.5 ชุดข้อมูล

ข้อมูลที่น่ามาใช้ในการงานวิจัยนี้ คือ ข้อมูลจากพาหนะและโทรศัพท์มือถือ (Vehicles and Mobile Probe Data) ซึ่งเป็นข้อมูลในช่วงเวลา 1 มกราคม 2563 ถึง 31 ธันวาคม 2563 ขนาดไฟล์ 90.7 GB จาก iTIC foundation เป็นข้อมูลการใช้งานของพาหนะที่ถูกรวบรวมจากระบบของ

พาหนะเอง และจากแอปพลิเคชันโทรศัพท์มือถือ โดยข้อดีของการใช้ข้อมูลนี้เมื่อเทียบกับการใช้เซ็นเซอร์ตรวจจับแบบเก่า (Fixed Sensors) ข้อมูลดังกล่าวจะมีความครอบคลุมพื้นที่มากกว่าและมีค่าใช้จ่ายที่ถูกกว่า ข้อมูลที่ได้มีความเที่ยงตรงมากกว่าการประมาณเวลาเดินทางจากความเร็วที่จับได้ขณะผ่านเซ็นเซอร์ ทั้งนี้ข้อมูลจากพาหนะ และโทรศัพท์มือถือประกอบไปด้วย 8 ฟีเจอร์ (Feature) ได้แก่

- VehicleID คือ ข้อมูลที่ใช้ระบุพาหนะแต่ละคันโดยพาหนะแต่ละคันใช้ VehicleID เดียวกันเสมอ ไม่มีการเปลี่ยนแปลง
- gpsvalid คือ ค่าที่บ่งบอกถึงสถานะของระบบตรวจจับข้อมูลพาหนะว่ามีสัญญาณหรือไม่ในขณะที่เก็บข้อมูล
- lat และ lon คือ ค่าละติจูดและลองจิจูดของพาหนะ
- timestamp คือ เวลาในขณะที่เก็บข้อมูล
- speed คือ ความเร็วของพาหนะในขณะที่เก็บข้อมูล
- for_hire_light คือ ค่าสำหรับพาหนะประเภทแท็กซี่ที่ใช้ในการบ่งบอกถึงสถานะไฟรับผู้โดยสารโดยจะมีค่าเป็น 0 และ 1 หมายถึงปิดไฟและเปิดไฟตามลำดับ
- engine_acc คือ ค่าที่บ่งบอกถึงสถานะเครื่องยนต์ของพาหนะโดยจะมีค่าเป็น 0 และ 1 หมายถึงเครื่องยนต์ไม่ทำงานและทำงานตามลำดับ

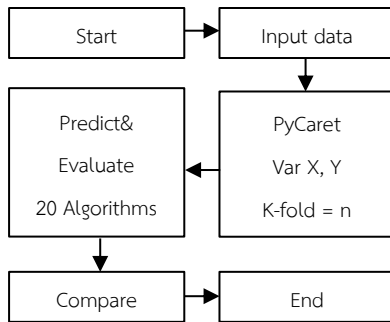
ตัวอย่างของข้อมูลดังกล่าวสามารถแสดงในรูปที่ 1

	VehicleID	gpsvalid	lat	lon	timestamp	speed	heading	for_hire_light	engine_acc
0	RsiQyWE3MJC3AgmVHLg5SL0	1	13.74835	100.65628	2019-12-31 23:59:25	0	10	0	1
1	UXuWwvC5gFygmX7AndpEnO8	1	13.81607	100.64795	2019-12-31 23:58:46	0	33	0	0
2	RdZ-aLutHgiwstAmBwICD3il	1	13.82811	100.63003	2020-01-01 00:00:12	30	188	1	1
3	aOm4w40QJAJUabOgRvQWu7rs	1	13.83277	100.57138	2019-12-31 23:59:58	47	30	1	1
4	ZEH01Tt6GRPK8VhuppKXOXng	1	13.74802	100.54831	2020-01-01 00:00:09	0	12	1	1

รูปที่ 1 ข้อมูลจากพาหนะ และโทรศัพท์มือถือ

3.6 ขั้นตอนการสร้างต้นแบบ และ เปรียบเทียบ

ชุดข้อมูลหลังจากผ่านการจัดการแล้วจะอยู่ในรูปแบบไฟล์ csv เพื่อทำการ input ข้อมูลลงในโปรแกรม จากนั้นจึง ประกาศฟิเจอร์ที่ต้องการคาดการณ์ (Y) และ ฟิเจอร์ที่ใช้สร้างต้นแบบ (X) แล้วใช้ PyCaret ในการแบ่งข้อมูลเพื่อทำ K-fold cross validation สร้างต้นแบบจากอัลกอริทึม และ ประเมินผล ทำซ้ำจนครบอัลกอริทึมทั้งหมด แผนผังขั้นตอนการสร้างต้นแบบ สามารถแสดงได้ดังรูปที่ 2



รูปที่ 2 แผนผังขั้นตอนการสร้างต้นแบบ

4. ผลการดำเนินงาน

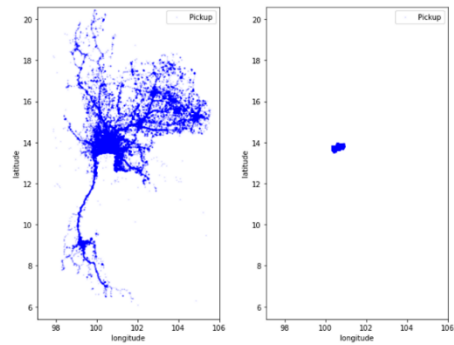
4.1 การจัดการชุดข้อมูล

ทำการแปลงชุดข้อมูลให้อยู่ในรูปแบบของ Origin-Destination pairs โดยเลือกเฉพาะรถแท็กซี่ซึ่งกะกำลังปิดไฟรับผู้โดยสาร คำนวณระยะเวลาเดินทาง การกระจัดระหว่างจุดเริ่มต้นและจุดสิ้นสุด แกะไขข้อมูลวันที่ออกมาเป็นตาราง เนื่องจากโปรแกรมไม่สามารถใช้วันที่ในรูปแบบนี้ได้ หลังจากแปลงชุดข้อมูลแล้วจะได้ข้อมูล ดังแสดงในรูปที่ 3

Pickuplat	Pickuplon	Droptlat	Droptlon	AverageSpeed	TripDuration	PickupSecOfDay	DayofWeek	DayofMonth	DroptSecOfDay	Month	Displacement	
0	13.76747	100.40770	13.83370	54.733333	1139	88370	2	1	1109	1	15.619978	
1	13.77579	100.40927	13.87485	100.40960	52.333333	719	32491	3	2	33210	1	11.239557
2	13.77460	100.41061	13.87527	100.40964	44.250000	900	30717	3	2	31817	1	11.060413
3	13.76179	100.41058	13.87490	100.40964	37.857143	1020	30228	5	4	31248	1	9.869506
4	13.77529	100.42872	13.88921	100.40621	46.533333	1140	42446	0	13	43588	1	11.989916

รูปที่ 3 ข้อมูล Origin-Destination pairs

งานวิจัยนี้จะเลือกใช้ข้อมูลทริปที่อยู่ในกรุงเทพมหานคร เนื่องจากจำนวนของตัวอย่างในจังหวัดอื่น ๆ ทั้งหมดรวมกันแล้วมีเพียง 30.43% จากทริปทั้งหมด นอกนั้นจะเป็นของกรุงเทพมหานครจังหวัดเดียว รูปที่ 4 แสดงกราฟที่พล็อตจุดเริ่มต้น ก่อน และ หลังลบทริปจังหวัดอื่น ๆ ออกจากชุดข้อมูล



รูปที่ 4 กราฟจุดเริ่มต้น ก่อน และ หลังลบทริปจังหวัดอื่น

4.2 ผลการทดลอง

เมื่อทำการ input ข้อมูลในโปรแกรมแล้ว ผู้วิจัยได้ทำการปรับพารามิเตอร์ PyCaret ของ K-fold ที่ 5 (ค่าเริ่มต้นในการแบ่งข้อมูล train และ test อยู่ที่ 70 และ 30) จากนั้นจึงเปิดการใช้การ์ดจอในการสร้างโมเดล (สำหรับบางอัลกอริทึมมีการใช้การ์ดจอเพื่อช่วยในการสร้างต้นแบบ)

ผู้วิจัยได้ใช้โปรแกรม Jupyter Notebook ในการรันการทดลองบนคอมพิวเตอร์

- CPU : AMD Ryzen 7 3700X 8-Core Processor
- Memory : 32 GB
- GPU : NVIDIA GeForce RTX 3060 Ti

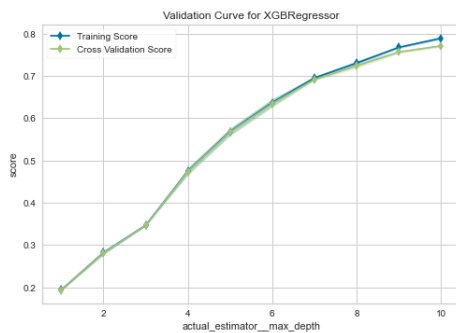
ซึ่งได้ผลสรุป อัลกอริทึม 5 อันดับ ที่ต้นแบบมีค่า RMSE น้อยที่สุด ดังแสดงในตารางที่ 1

ตารางที่ 1 ผลการทดลอง

Model	RMSE	Training time (Sec)
Extreme Gradient Boosting	163.3069	14.5100
Random Forest Regressor	169.4961	944.1080
Extra Trees Regressor	194.6996	702.2720
Light Gradient Boosting Machine	198.2688	5.1040
CatBoost Regressor	214.0307	13.4420

4.3 อภิปรายผลการทดลอง

จากตารางผลการทดลองจะพบว่าอัลกอริทึม XGBoost (Extreme Gradient Boosting) มีค่า RMSE น้อยที่สุด และใช้เวลาในการสร้างต้นแบบที่น้อยมากอีกด้วย เมื่อทำการวิเคราะห์ต้นแบบนี้เพิ่มเติมจะพบว่า Training Score (คะแนนความแม่นยำในการคาดการณ์ของต้นแบบเทียบกับข้อมูลชุด train) กับ Cross Validation Score (คะแนนความแม่นยำในการคาดการณ์ของต้นแบบเทียบกับข้อมูลชุด test) มีค่าที่สูง และใกล้เคียงกัน หมายความว่าต้นแบบนี้ไม่ประสบปัญหา overfitting ดังแสดงในรูปที่ 5



รูปที่ 5 Validation Curve ของ XGBoost

จากผลข้างต้น อาจสรุปได้ว่า XGBoost เป็นอัลกอริทึมที่สามารถนำไปใช้งานได้ ด้วยประสิทธิภาพต้นแบบที่ดี และเวลาสร้างที่ค่อนข้างน้อย

อย่างไรก็ตามต้นแบบจากอัลกอริทึมทั้ง 5 อันตั้บนี้ยังสามารถพัฒนาประสิทธิภาพได้อีก ด้วยวิธีปรับปรุงพารามิเตอร์ และสร้างต้นแบบใหม่ซ้ำ ๆ จนได้ต้นแบบที่มีประสิทธิภาพสูงขึ้น ซึ่งอาจจะทำให้อัลกอริทึมที่ใช้เวลาในการสร้างต้นแบบน้อยมีความได้เปรียบ ดังนั้นจึงสรุปได้ว่าจากผลการทดลองมีอัลกอริทึม 3 ตัวที่มีศักยภาพในการพัฒนาต่อได้แก่ XGBoost LightGBM และ CatBoost ส่วนของ Random Forest Regressor และ Extra Trees Regressor นั้นใช้เวลาในการสร้างต้นแบบมาก แม้จะมี RMSE ที่น้อยกว่า LightGBM และ CatBoost ก็ตาม แต่จะเสียเปรียบในการนำไปพัฒนาปรับปรุงพารามิเตอร์

5. สรุป

ในงานวิจัยนี้ได้นำเสนอการสร้างต้นแบบในการคาดการณ์เวลาเดินทางบนท้องถนนของกรุงเทพมหานคร โดยใช้ข้อมูลพาหนะและโทรศัพท์มือถือจาก iTIC foundation ซึ่งเป็นข้อมูลที่สามารถเข้าถึงได้แบบไม่มีค่าใช้จ่าย โดยใช้วิธีการเรียนรู้ของเครื่องในการสร้างต้นแบบ จากการเปรียบเทียบประสิทธิภาพของต้นแบบที่สร้างโดยอัลกอริทึมทั้งหมด 20 อัลกอริทึม จะมีอัลกอริทึมที่สามารถสร้างต้นแบบที่มีประสิทธิภาพเหมาะสมในการนำไปใช้งาน หรือพัฒนาต่อ 3 ตัว ได้แก่ XGBoost LightGBM และ CatBoost

ข้อจำกัดในด้านข้อมูล อาจส่งผลกับงานวิจัยนี้เนื่องจาก iTIC มีการเก็บข้อมูลจากรถแท็กซี่เพียง 4,634 คันเท่านั้น และจากการระบาดของโรคติดเชื้อไวรัสโคโรนา ส่งผลให้การจราจรอยู่ในสภาพไม่ปกติ จากการ Lock Down, Work from Home, Curfew สิ่งเหล่านี้อาจส่งผลกระทบต่อข้อมูลปกติ และส่งผลกระทบต่อผลการสร้างต้นแบบ

ในอนาคตอาจมีการนำชุดข้อมูลที่ส่งผลกับการจราจรเข้ามาใช้เพิ่มเติมในการสร้างต้นแบบ เพื่อให้ต้นแบบมีประสิทธิภาพในการคาดการณ์มากขึ้น เช่น ข้อมูลสภาพอากาศ ไฟจราจร ระยะทางจริง จากโครงข่ายถนน ความหนาแน่นของการจราจรในแต่ละเขต หรือการใช้ข้อมูล Origin-Destination pair จากแหล่งข้อมูลอื่นมาพัฒนาต่อยอด

6. เอกสารอ้างอิง

- [1] F. Li, Z-P. Fan, B-B. Cao and X. Li, "Logistics service mode selection for last mile delivery: an analysis method considering customer utility and delivery service cost," *Sustainability*, vol. 13, no. 1, p. 284, 2021.
- [2] R. Gevaers, E. Van de Voorde and T. Vanelander, "Characteristics of innovations in last-mile logistics-using best practices, case studies and making the link with green and

- sustainable logistics,” *Association for European Transport and contributors.*, pp. 1-21, 2009.
- [3] K. K. Boyer, A. M. Prud'homme and W. Chung, “The last mile challenge: evaluating the effects of customer density and delivery window patterns,” *Journal of business logistics.*, vol. 30, no. 1, pp. 185-201, 2009.
- [4] J. D. Cortes and Y. Suzuki, “Last-mile delivery efficiency: en route transloading in the parcel delivery industry,” *International Journal of Production Research.*, pp. 1-18, 2021.
- [5] MARKETINGOOPS!, “3 กลยุทธ์ ‘ธุรกิจขนส่งพัสดุ’ รับมือสงครามราคา ปริมาณส่ง 4 ล้านชิ้นต่อวัน มูลค่า 6.6 หมื่นล้าน,” [ออนไลน์]. แหล่งที่มา: <https://www.marketingoops.com/reports/industry-insight/transport-and-logistics-strategy-2020/>. [วันที่เข้าถึง 11 พฤษภาคม 2564].
- [6] Y. Geng, E. Liu, R. Wang and Y. Liu, “Deep Reinforcement Learning Based Dynamic Route Planning for Minimizing Travel Time,” in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2021.
- [7] N. Frohner, M. Horn and G. R. Raidl, “Route duration prediction in a stochastic and dynamic vehicle routing problem with short delivery deadlines,” *Procedia Computer Science.*, vol. 180, pp. 366-370, 2021.
- [8] TomTom, “TomTom Traffic Index ranking 2020,” [Online]. Available: https://www.tomtom.com/en_gb/traffic-index/ranking/. [Accessed 11 May 2021].
- [9] G. Yu and Y. Yang, “Dynamic routing with real-time traffic information,” *Operational Research.*, vol. 19, no. 4, pp. 1033-1058, 2019.
- [10] S. Parinya, “Developing applications for vehicle routing problems with real time data acquisition,” *International Journal of Simulation - Systems, Science & Technology.*, vol. 20, no. 2, pp. 1-6, 2019.
- [11] Hadiyanto, W. Budi, Maryono, E. Subowo, E. Sediyo and Farikhin, “Ant Colony Algorithm for Determining Dynamic Travel Routes Based on Traffic Information from Twitter,” in *The 4th International Conference on Energy, Environment, Epidemiology and Information System (ICENIS 2019)*, 2019.
- [12] L. Huang and L. Xu, “Research on Taxi Travel Time Prediction Based on GBDT Machine Learning Method,” in *2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)*, 2018.
- [13] B. Deb, S. R. Khan, K. T. Hasan, A. H. Khan and M. A. Alam, “Travel Time Prediction using Machine Learning and Weather Impact on Traffic Conditions,” in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, 2019.
- [14] D. Lin, W. Tsao, C. Yu, H. Liu and Y. Chang, “The Travel Time Prediction by Machine Learning Methods with Traffic Data in Chiayi City, Taiwan,” in *4th International Conference on Electromechanical Control Technology and Transportation (ICECTT)*, 2019.
- [15] M. Mohri, A. Rostamizadeh and A. Talwalkar, *Foundations of machine learning*, MIT press, 2018.
- [16] T. Dietterich, “Overfitting and undercomputing in machine learning,” *ACM computing surveys (CSUR).*, vol. 27, no. 3, pp. 326-327, 1995.
- [17] J. D. Rodriguez, A. Perez and J. A. Lozano, “Sensitivity analysis of k-fold cross validation in prediction error estimation,” *IEEE transactions on pattern analysis and machine intelligence.*, vol. 32, no. 3, pp. 569-575, 2009.

- [18] S. Oh, Y.-J. Byon, K. Jang and H. Yeo, "Short-term travel-time prediction on highway: a review of the data-driven approach," *Transport Reviews.*, vol. 35, no. 1, pp. 4-32, 2015.
- [19] B. Qiu and W. Fan, "Travel time forecasting on a freeway corridor: a dynamic information fusion model based on the random forests approach," *Smart and Resilient Transportation.*, vol. 3, no. 2, pp. 131-148, 2021.
- [20] J. W. C. Lint, "Reliable real-time framework for short-term freeway travel time prediction," *Journal of Transportation Engineering.*, vol. 132, 2006.