การสร้างโมเดลเพื่อทำนายโอกาสการเกิดความล่าช้าของเที่ยวบินโดยใช้เทคนิคการเรียนรู้ของเครื่องประเภท Boosting

Building the Predicting model of Flight Delay Using Boosting Algorithm

พนิดา กาศกลางดอน
สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์
มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร
กรุงเทพมหานคร, ประเทศไทย
E-mail:panida.joyce@g.swu.ac.th

***บทคัดย่อ*—งานวิจัยนี้จัดทำมาเพื่อนำเสนอการประยุกต์เทคนิคการเรียนรู้ของเครื่องเพื่อสร้างโมเดลในการทำนายโอกาสการเกิดความล่าช้าของเที่ยวบิน จากการใช้ข้อมูลเที่ยวบินของประเทศสหรัฐอเมริกาในปี 2015 โดยมีจุดประสงค์เพื่อวิเคราะห์ข้อมูลเพื่อหาปัจจัยสำคัญที่ส่งผลกระทบให้เกิดความล่าช้าต่อเที่ยวบินอย่างมีนัยสำคัญแล้วนำมาสร้างเป็นข้อมูลใหม่ เพื่อช่วยให้การเรียนรู้ของเครื่องสามารถนำข้อมูลไปเรียนรู้และสามารถสร้างโมเดลที่ได้ค่าความถูกต้องที่สูงสุดในการทำนายเที่ยวบินที่จะเกิดความล่าช้า โดยผลในการทำนายจากการใช้โมเดลทำนาย ทั้ง 5 โมเดล 1.Gradient Boosting 2.CatBoost 3.SVM 4.Logistic Regression 5.Random forest พบว่าค่าความถูกต้องของโมเดลที่ใช้เทคนิค catboosting มีประสิทธิภาพการทำนายสูงสุด มีค่าร้อยละ 86.8 โดยการทำนายนี้จะสามารถช่วยให้องค์กรการขนส่งทางอากาศ จัดการตารางการเดินทางของสายการบินจะสามารถปรับเที่ยวเที่ยวบินหรือพัฒนา ระยะเวลาที่เหมาะสมในแต่ละเที่ยวบิน**

***ABSTRACT*—This research was prepared to present the application of machine learning techniques to create models to predict the likelihood of flight delays. From the use of flight data from the United States in 2015 with the aim of analyzing the data to find important factors that cause significant delays on flights and create new data To help the machine learn to use the information to learn and to create models that provide the highest accuracy in predicting flight delays The prediction results from using 5 predictive models 1.Gradient Boosting 2.CatBoost 3.SVM 4.Logistic Regression 5.Random forest found that the accuracy of the models using catboosting techniques has the highest predictive efficiency. 86.8 percent of these predictions will be able to help air transportation organizations Manage the airline's travel schedule to be able to adjust flights or develop Appropriate duration for each flight**

คำสำคัญ—ความล่าช้าของเที่ยวบิน, FlightDelay, GradintsBoosting, Timedelay, catboosting, Randon forest, Logistic regression, SVM

1. บทนำ

เนื่องจากปัจจุบันการเดินทางระยะไกลส่วนใหญ่คือการเดินทางโดยเครื่องบินซึ่งมีค่าใช้จ่ายค่อนข้างสูงแต่ก็สามารถเดินทางได้อย่างรวดเร็วหลายเท่าตัวเมื่อเทียบกับการเดินทางของทางบกและทางน้ำ นั้นจึงหมายความว่าผู้โดยสารนั้นให้ความสำคัญของเวลาเป็นอย่างมาก ซึ่งผู้ให้บริการสายการบินนั้นจำเป็นต้องมีความตรงต่อเวลา ปัจจุบันเครื่องบินดีเลย์นับเป็นปัญหาใหญ่ของการขนส่งทางอากาศ โดยเฉพาะในช่วงเทศกาล ซึ่งก่อให้เกิดปัญหาความไม่สะดวกต่อผู้โดยสารเป็นอย่างมาก โดยเฉพาะในสหรัฐฯ ในปีค.ศ. 2002 ปัญหาเครื่องบินดีเลย์ หรือการที่เครื่องบินบินถึงจุดหมายปลายทางล่าช้าเกินกว่า 15 นาที คิดเป็นสัดส่วนเกือบ 1 ใน 4 ของเที่ยวบินทั้งหมด และส่งผลกระทบทางตรงต่อต้นทุนของสายการบินของสหรัฐฯ มากถึงปีละ 200,000 ล้านบาท ท่าอากาศยานซึ่งมีรันเวย์จำกัด ทำให้เครื่องบินต้องต่อคิวหลายลำกว่าจะบินขึ้นได้ หากเที่ยวบินใดเที่ยวบินหนึ่งดีเลย์ จะส่งผลกระทบต่อเนื่องถึงเที่ยวบินอื่นๆ ของที่ต่อจากเครื่องบินลำนั้นต้องดีเลย์ตามไปด้วย ความล่าช้าของเครื่องบินเกิดได้จากหลายสาเหตุ เช่น ความล่าช้าก่อนออกเดินทาง ความล่าช้าระหว่างขับเคลื่อนเพื่อขึ้นบิน ความล่าช้าระหว่างเส้นทางการบิน และความล่าระหว่างขับเคลื่อนมาถึงหลุมจอด ซึ่งความล่าช้าเหล่านี้อาจจะเป็นความล่าช้าสะสม หรือต่อเนื่องมาจากกิจกรรมก่อนหน้าที่เกิดขึ้น โดยจะสามารถวัดได้จากการเปรียบเทียบเวลาที่กำหนดเอาไว้กับเวลาที่วัดได้จริง ดังนั้นสาเหตุความล่าที่เกิดขึ้นอาจจะเกิดจากการบริหารจัดการจราจรทางอากาศ ซึ่งจะทำการพิจารณาดูข้อมูลของเที่ยวบินและข้อจำกัดต่างๆ เพื่อทำการจัดการช่วงเวลาที่เหมาะสมสำหรับการออกเดินทางของทางอากาศยาน

ในกรณีของประเทศไทยที่มีรายได้กว่า 1.29 ล้านล้านบาทต่อปีมาจากการท่องเที่ยวซึ่งในไทยพบว่ามีนักท่องเที่ยวจำนวนกว่า 26.5 ล้านคน สิ่งอำนวยความสะดวกและเป็นประตูแรกที่ต้อนรับเข้าสู่ประเทศไทยก็คือท่าอากาศยาน การเกิดความล่าช้าตั้งแต่เริ่มต้นการท่องเที่ยวอาจส่งผลกระทบต่อภาพลักษณ์ของประเทศไทยและอาจนำไปสู่ปัญหาที่ทำให้ยอดนักท่องเที่ยวลดลง ผู้จัดทำจึงเล็งเห็นความสำคัญของการแก้ไขปัญหาความล่าช้าของเที่ยวบินโดยใช้ชุดข้อมูลเที่ยวบินของประเทศสหรัฐฯในปี 2015 ที่มีเผยแพร่ทางสาธารณะใน dataset บนเว็ป kaggle มาเป็นกรณีศึกษาเพื่อใช้เป็นแนวทางในการพัฒนาระบบการเรียนรู้ของเครื่องที่อาจเป็นประโยชน์ต่อการแก้ไขปัญหาความล่าช้าของเที่ยวบินในประเทศไทย

จากปัญหาความล่าช้าของเที่ยวบินดังกล่าว ทางผู้จัดทำจึงทำการวิเคราะห์ข้อมูลโดยการพลอตกราฟเพื่อหาความสัมพันธ์ของข้อมูลว่าข้อมูลชนิดใดมีความสัมพันธ์กับความล่าช้ามากที่สุด จึงได้มีการนำเทคนิคการสร้างโมเดลการแก้ปัญหาประเภท classification เพื่อการทำนายโอกาสที่จะเกิดความล่าช้าของเที่ยวบิน โดยโมเดลที่จัดทำขึ้นจะช่วยให้ทางอากาศยานสามารถบริหารจัดการตารางเที่ยวบินได้อย่างเหมาะสมและช่วยเพิ่มการกระจายตัวของเที่ยวบินเพื่อหลีกเลี่ยงหรือลดความล่าช้าสะสมจากกิจกรรมก่อนหน้าที่เกิดขึ้น ซึ่งอาจช่วยให้สายการบินเกิดความเสียหายน้อยที่สุด ที่มีสาเหตุจากความล่าช้าของเที่ยวบิน และส่งผลให้ผู้โดยสารสามารถเดินทางถึงจุดหมายตามตารางเวลาที่กำหนดไว้

2. วัตถุประสงค์

2.1 เพื่อศึกษาข้อมูลและนำมาใช้วิเคราะห์ปัจจัยที่ผลให้เกิดความล่าช้าของเที่ยวบิน

2.2 เพื่อสร้างโมเดลการทำนาย โดยคัดเลือกเทคนิคที่ให้ประสิทธิภาพของความแม่นยำมากที่สุด ตัวอย่างเทคนิคที่นำมาใช้เช่น Gradient Boosting, CatBoost, SVM, Logistic Regression, Random forest

2.3 เพื่อทำนายหาโอกาสการเกิดความล่าช้าในแต่ละเที่ยวของเที่ยวบิน

3. ขอบเขตงานวิจัย

3.1 ที่มาข้อมูลสำนักงานสถิติการขนส่งแห่งสหรัฐอเมริกา (DOT) ของกระทรวงคมนาคมติดตามประสิทธิภาพการบินตรงเวลาของเที่ยวบิน ในชุดข้อมูลขการล่าช้าของเที่ยวบินและการยกเลิกเที่ยวบินในปี 2558 ที่ถูกนำมาเผยแพร่ต่อบน dataset บนเว็ปไซร์ kaggle

3.2 การวิเคราะห์ปัจจัยที่ส่งผลให้เกิดความล่าช้าของเที่ยวบินมีจำนวน 12 แอตทริบิวต์ได้แก่ ปีที่เดินทาง (YEAR) เดือนที่เดินทาง (MONTH) วันที่เดินทาง(DAY) วันที่เดินทางในสัปดาห์(DAY_OF_WEEK) เวลาออกเดินทางตามแผน (SCHEDULED_DEPARTURE) เวลาที่เครื่องออกเดินทางจริง (DEPARTURE_TIME) ความล่าช้าที่เครื่องออกเดินทางจริงกับกำหนดการ(DEPARTURE_DELAY) ระยะเวลาที่ใช้ระหว่างเกตไปกระทั่งเวลาที่ล้อเครื่องถูกเก็บ (TAXI_OUT) ระยะเวลาที่ใช้ตามแผน(SCHEDULED_TIME) เวลาที่ล้อถูกเก็บขึ้น (WHEELS_OFF) เวลาถึงที่หมายตามแผน (SCHEDULED_ARRIVAL) ระยะทางการเดินทาง (DISTANCE)

3.3 วิธีการเรียนรู้แบบ classification จะใช้เทคนิคการจำแนกประเภทเพื่อจำแนกโอกาสเกิดความล่าช้า ของเที่ยวบินว่ามีโอกาสหรือไม่ เทคนิคการจำแนก ประเภทข้อมูลเป็นเทคนิคหนึ่งที่สำคัญของการ สืบค้นความรู้บนฐานข้อมูลขนาดใหญ่หรือดาต้าไมน์ นิง เทคนิคการจำแนกประเภทข้อมูลเป็นกระบวนการ สร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมา ให้จากกลุ่มตัวอย่างข้อมูลที่เรียกว่าข้อมูลสอนระบบ ที่แต่ละแถวของข้อมูลประกอบด้วยฟิลด์หรือแอทริบิวท์จำนวนมาก

3.4 การหาค่าความถูกต้องของโมเดลจะประเมินจากค่าความถูกต้อง(Accurency)

4. แนวคิดและทฤษฎีที่เกี่ยวข้อง

4.1 Pandas Dataframe เป็นโครงสร้างข้อมูลแบบตารางสองมิติที่ไม่แน่นอนซึ่งมีความแตกต่างกันโดยมีแกนข้อความ (แถวและคอลัมน์) กรอบข้อมูลเป็นโครงสร้างข้อมูลแบบสองมิติเช่นข้อมูลถูกจัดแนวแบบตารางในแถวและคอลัมน์ Pandas DataFrame ประกอบด้วยสามองค์ประกอบหลักคือข้อมูลแถวและคอลัมน์

4.2 อัลกอริทึม Decision Tree แผนผังการตัดสินใจเป็นเครื่องมือสนับสนุนการตัดสินใจที่ใช้กราฟหรือรูปแบบของการตัดสินใจและผลที่เป็นไปได้รวมถึงผลลัพธ์เหตุการณ์โอกาสต้นทุนทรัพยากรและยูทิลิตี้ มันเป็นวิธีหนึ่งในการแสดงอัลกอริทึมที่มีเพียงคำสั่งควบคุมตามเงื่อนไข

4.3 อัลกอริทึม Gradients Boosting เป็นเทคนิคการเรียนรู้ของเครื่องสำหรับ=ปัญหาregressionและclassificationซึ่งสร้างแบบจำลองการทำนายในรูปแบบของชุดweak predictionทั้งหมดซึ่งโดยทั่วไปจะเป็น decision tree มันสร้างโมเดลในรูปแบบที่ชาญฉลาดเช่นเดียวกับวิธีการboostingแบบอื่น ๆ และสามารถคำนวณค่า loss function ของแต่ละโมเดลได้ในตัวโมเดลเองเพื่อทำการ optimization

4.4 อัลกอริทึม Logistic Regression

การวิเคราะห์การถดถอยโลจิสติค มีวัตถุประสงค์เพื่อศึกษาว่าตัวแปรอิสระหรือตัวแปรทำนายใดบ้างที่สามารถอธิบายตัวแปรเกณฑ์ (ตัวแปรตาม) ซึ่งเป็นตัวแปรทวิหรือตัวแปรพหุกลุ่ม Analysis จะใช้ลักษณะหรือธรรมชาติของตัวแปรตอบสนอง (Response) เป็นตัวกำหนด ตัวแปรอิสระใดบ้างที่สามารถใช้อธิบายโอกาสการเกิดเหตุการณ์หรือการไม่เกิดเหตุการณ์ ที่สนใจตามตัวแปรตามหรือตัวแปรเกณฑ์ พร้อม ทั้งศึกษาระดับความสัมพันธ์ของตัวแปรทำนาย แต่ละตัว เพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่ สนใจ จากสมการโลจิสติคที่เหมาะสม โดยเลือก ตัวแปรที่เหมาะสมเพื่อทำให้เปอร์เซ็นต์ของความ ถูกต้องในการทำนายมีค่าสูงสุด

โมเดลทางคณิตศาสตร์

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

4.5 อัลกอริทึม Support Vector Machine (SVM) เป็นตัวจำแนกเชิงเส้น (Linear Classifier) แบบ 2 คลาส ซึ่งเป็นที่ยอมรับถึงประสิทธิภาพของการจำแนกที่เหนือว่าวิธีการจำแนกอื่นๆ ข้อได้เปรียบของ SVM คือมีประสิทธิภาพในการจำแนกข้อมูลที่มีมิติจำนวนมากได้ นอกจากนี้การใช้ฟังก์ชันเคอร์เนล (Kernel Function) เพื่อแปลงข้อมูลไปยังมิติที่สูงขึ้นในปริภูมิคุณลักษณะ(Feature Space)สามารถจำแนกข้อมูลที่มีความคลุมเครือได้อย่างมีประสิทธิภาพ หลักการของSVM คือการหาเส้นตรงที่มีมาร์จินที่โตที่สุด (Maximum Margin) ที่สามารถแบ่งข้อมูลออกเป็น 2 คลาส การใช้เส้นตรงสำหรับแบ่งข้อมูลเป็น 2 กลุ่มด้วยมาร์จินที่โตที่สุด (Maximum Margin)

เป็นวิธีที่การันตีได้ว่าจะสามารถแยกข้อมูลได้โดยมีความผิดพลาดน้อยที่สุด

4.6 อัลกอริทึมแบบ CatBoost เป็นอัลกอริทึมการเรียนรู้ของเครื่องที่ใช้ gradient boosting บน decision trees CatBoost มีความยืดหยุ่นในการให้ดัชนีของคอลัมน์หมวดหมู่เพื่อให้สามารถทำOneHotEncodingโดยใช้ one_hot_max_size (ใช้one-hot encodingสำหรับข้อมูลที่เป็น Class ทั้งหมดที่มีจำนวนค่าแตกต่างกันน้อยกว่าหรือเท่ากับ ค่าพารามิเตอร์ที่กำหนด) ไม่จำเป็นจะต้องใช้ข้อมูลจำนวนมากเพื่อให้ได้ผลลัพท์ที่ดี

5. งานวิจัยที่เกี่ยวข้อง

(Karthik Gopalakrishnan and Hamsa Balakrishnan,2017) ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการต่างๆในการทำนายความล่าช้าในเครือข่ายการขนส่งทางอากาศและ พิจารณาโมเดลสามแบบ: โมเดลรวมที่พัฒนาล่าสุดของความล่าช้าของเครือข่ายการเคลื่อนที่ซึ่งเราจะเรียกว่า Markov Jump Linear System (MJLS), เทคนิคการเรียนรู้ของเครื่องจักรแบบ classifacation เช่น Classi และ Reg ต้นไม้ (CART) และสามผู้สมัครสถาปัตยกรรมเครือข่าย Neural Neural (ANN) ซึ่งแสดงให้เห็นว่าประสิทธิภาพการทำนายอาจแตกต่างกันอย่างมีนัยสำคัญขึ้นอยู่กับการเลือกโมเดล / อัลกอริทึมและประเภทของการทำนาย (ตัวอย่าง เช่น classification กับ regression) นอกจากนี้ยังพูดคุยถึงความสำคัญของการเลือกตัวแปรทำนายหรือคุณสมบัติที่เหมาะสมเพื่อปรับปรุงประสิทธิภาพของอัลกอริทึมเหล่านี้ โมเดลได้รับการประเมินโดยใช้ข้อมูลการปฏิบัติงานจาก National Airspace System (NAS) ของสหรัฐอเมริกา ANN แสดงให้เห็นว่าเป็นอัลกอริธึมที่ดี สำหรับปัญหาการจำแนกประเภทซึ่งมีความแม่นยำโดยเฉลี่ยอยู่ที่ 94% ในการทำนายความล่าช้าจาก 100 การเชื่อมโยงที่ล่าช้าที่สุดจะเกิน 60 นาทีหรือสองชั่วโมงในอนาคต อย่างไรก็ตามโมเดล MJLS นั้นดีกว่าคาดการณ์ การเชื่อมโยงที่ผิดปกติและมีข้อผิดพลาดในการทำนายค่าเฉลี่ย 4.7 นาที สำหรับปัญหาการถดถอยเป็นเวลา 2 ชั่วโมง โมเดล MJLS ยังดีกว่าในการทำนายความล่าช้าขาออกที่สนามบินหลัก 30 แห่งโดยมีข้อผิดพลาดเฉลี่ย 6.8 นาทีเป็นเวลา 2 ชั่วโมง ผลของปัจจัยชั่วคราวและการกระจายตัวเชิงพื้นที่ของความล่าช้าในปัจจุบันในการทำนายความล่าช้าในอนาคตนอกจากนี้ยังมีการเปรียบเทียบ โมเดล MJLS ซึ่งได้รับการออกแบบมาเป็นพิเศษเพื่อดักจับการเคลื่อนที่ของการเคลื่อนที่ของอากาศ โดยรวมใช้ประโยชน์จากปัจจัยเหล่านี้และมีประสิทธิภาพเหนือกว่า ANN ในการทำนายการกระจายของความล่าช้าเชิงพื้นที่ในอนาคต ในลักษณะนี้การแลกเปลี่ยนระหว่างความเรียบง่ายของแบบจำลองกับความแม่นยำในการทำนายถูกเปิดเผย

(Roshni Musaddi 1, Anny Jaiswal 2, Pooja J 3, Mansvi Girdonia 4, Minu M.S 5, 2018) ได้มีการศึกษาใช้อัลกอริทึมเพื่อคาดการณ์ความล่าช้าของเที่ยวบิน โดยใช้ Python ใน Visual Studio Code และใช้การจำแนกแบบไบนารีเพื่อเตรียมแบบจำลองที่สามารถทำนายความล่าช้าได้ข้อกำหนดตามดัชนี - Binary Classification, Visual Studio Code, Regularities, Python, Transportation, Complexities, Air Traffic Flow Managemen ในขั้นตอนแรกจะทำการทำความสะอาดชุดข้อมูล เช่น การล้างวันที่และเวลาและลบข้อมูลที่ไม่จำเป็นออก โดยทั่วไปเราจะเปรียบเทียบสายการบินตามคำอธิบายทางสถิติของสายการบินอื่นและทำการกระจายล่าช้า โดยการจัดอันดับสายการบิน การใช้อัลกอริทึมกับ Visual Studio Code ใน python จะดึงข้อมูลที่ต้องการและปรับแต่งชุดข้อมูล โดยการลบข้อมูลที่ไม่จำเป็นชุดข้อมูลที่ถูกแก้ไขนี้จะถูกแปลงเป็น sparse matrix ซึ่งใช้ Grid search บนโมเดล Random Forest เพื่อให้ได้ ROC Curve ผลลัพธ์จะถูกจัดเก็บเป็นสองกลุ่มโดย 0 เมื่อแสดงว่าเที่ยวบินตรงเวลาและ 1 เมื่อเที่ยวบินล่าช้า โดยจะใช้ฮิสโทแกรมต่างๆเพื่อเปรียบเทียบสายการบินต่างๆเกี่ยวกับวันและสัปดาห์และเพื่อทราบว่าสายการบินใดที่ให้บริการดีที่สุดในแง่ของความล่าช้าน้อยกว่า โดยสรุปความล่าช้าเที่ยวบินให้ตัวเลือกต่าง ๆ แก่ผู้โดยสารก่อนที่พวกเขาเดินทางผ่านสายการบินเหล่านี้

6. วิธีการดำเนินงานวิจัย

6.1 กระบวนการสำรวจและวิเคราะห์ข้อมูล (Exploration data analysis)

การสำรวจข้อมูลเป็นขั้นตอนเริ่มต้นที่สำคัญในการในการประเมินชุดข้อมูลสำหรับการวิเคราะห์ข้อมูลในขั้นตอนสุดท้ายเนื่องจากข้อมูลมีปริมาณมาก ข้อมูลจะกอบไปด้วยข้อมูลที่มีความสอดคล้องและไม่สอดคล้อง หรือ ข้อมูลที่มีคามผิดพลาดขาดหายไป ดังนั้นจึงจำเป็นต้องจัดการกับข้อมูลเหล่านี้ และ สืบค้นข้อมูลที่มีความจำเป็นมาใช้ในการวิเคราะห์เพื่อให้ให้เข้าถึงข้อมูลย่อยและทราบถึงความสัมพันธ์ระหว่างข้อมูลต่างๆได้ ผลลัพธ์จากการสำรวจข้อมูลจะสามารถสรุปและแสดงในรูปแบบที่เป็นกราฟ ข้อมูลที่ใช้จะเป็นข้อมูลที่จะได้รับตั้งแต่เครื่องบินจอดที่ท่าอากาศยานจนถึงกระทั่งเครื่องบินเก็บล้อเพื่อลอยตัวจากรันเวย์

ตารางที่1 อธิบายรายละเอียดจำนวนข้อมูลที่ใช้ในการสำรวจ

Flight Dataset	ตัวอย่าง	แอตทริบิวต์	เที่ยวบินที่ไม่เกิดความล่าช้า	เที่ยวบินที่เกิดความล่าช้า
จำนวน	56,375	12	35,873	20,502

6.1.1 การคัดเลือกแอตทริบิวต์ (Select Feature) ที่มีความเกี่ยวข้องหรือเป็นปัจจัยที่ส่งผลต่อการทำนายความล่าช้า แอตทริบิวต์ที่ถูกคัดเลือกจะแสดงดังตารางที่ 2

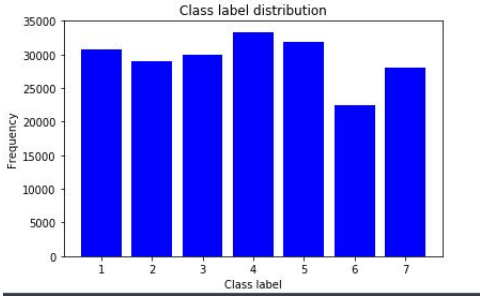
ตารางที่2 อธิบายแอตทริบิวต์ที่ใช้ในการสำรวจและวิเคราะห์ข้อมูล

ชื่อแอตทริบิวต์	คำอธิบาย
YEAR	ปีที่เดินทาง
MONTH	เดือนที่เดินทาง
DAY	วันที่เดินทาง
DAY_OF_WEEK	วันที่เดินทางในสัปดาห์
SCHEDULED_DEPARTURE	เวลาออกเดินทางตามแผน
DEPARTURE_TIME	เวลาที่เครื่องออกเดินทางจริง
DEPARTURE_DELAY	ความล่าช้าที่เครื่องออกเดินทางจริงกับกำหนดการ
TAXI_OUT	ระยะเวลาที่ใช้ระหว่างเกตไปกระทั่งเวลาที่ล้อเครื่องถูกเก็บ
WHEELS_OFF	เวลาที่ล้อถูกเก็บขึ้น
SCHEDULED_TIME	เวลาถึงที่หมายตามแผน
DISTANCE	ระยะทางการเดินทาง
SCHEDULED_ARRIVAL	เวลาถึงที่หมายตามแผน

6.1.2 Data visualization

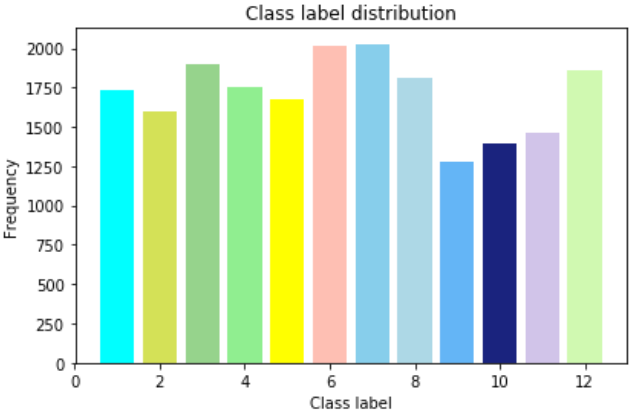
การนำข้อมูลมาสร้างกราฟเพื่อแสดงความสัมพันธ์หรือความแตกต่างของข้อมูลที่อาจส่งผลต่อการเรียนรู้ของเครื่อง

1. กราฟแสดงจำนวนเที่ยวบินที่เกิดความล่าช้ากับแอตทริบิวต์ DAY_OF_WEEK จากแผนภูมิกราฟแท่งรูปที่1 จะเห็นได้ว่าข้อมูลมีความแตกต่างกันอย่างชัดเจน ในวันที่6จะมีจำนวนเที่ยวบินที่เกิดความล่าช้าต่ำและวันที่1,2,3,7 จะมีจำนวนใกล้เคียงกันแล้ววันที่4กับ5ของสัปดาห์จะมีจำนวนเที่ยวบินที่เกิดความล่าช้าค่อนข้างสูง



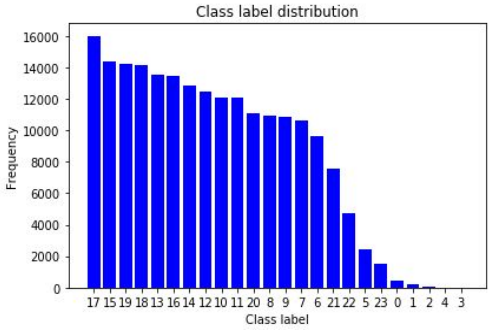
รูปที่1 แสดงจำนวนเที่ยวบินที่เกิดความล่าช้าในแต่ละวันของสัปดาห์

2. กราฟแสดงจำนวนเที่ยวบินที่เกิดความล่าช้ากับแอตทริบิวต์ MONTH จากแผนภูมิกราฟแท่งรูปที่2 จะเห็นได้ว่าช่วงตั้งแต่เดือน9 ถึง เดือน11 จะเกิดความล่าช้าค่อนข้างน้อยเมื่อเทียบกับเดือนอื่นๆ และในเดือนที่2กับเดือนที่4,5 จะมีจำนวนเที่ยวบินที่เกิดความล่าช้าน้อยรองลงมา แล้วเดือนที่เหลือจะมีจำนวนเที่ยวบินที่เกิดความล่าช้าพอๆกัน



รูปที่2 แสดงจำนวนเที่ยวบินที่เกิดความล่าช้าในแต่ละเดือน

3. สร้างแอตทริบิวต์ hour ที่นำข้อมูลจากแอตทริบิวต์ SCHEDULED_DEPARTURE โดยกำหนดให้แบ่งข้อมูลเป็นแต่ละช่วงของชั่วโมง โดยกราฟจะแสดงจำนวนเที่ยวบินที่เกิดความล่าช้าในแต่ละชั่วโมงจากกราฟแผนภูมิแท่งรูปที่3 จะพบว่า ชั่วโมงที่มีความล่าช้าต่ำจะอยู่ในชั่วโมงที่ 22, 5, 23, 0, 1, 2 ,3, 4 แล้วมีแนวโน้มที่จะเพิ่มขึ้นในชั่วโมงที่ 8, 20, 9, 7,6,21 แล้วมีแน้วโน้มที่สูงขึ้นอีกระดับนึงในชั่วโมงที่ 14, 12, 11, 10 แล้วขึ้นสูงต่อในชั่วโมงที่ 17,15,19,18,16,13 ซึ่งค่อนข้างเห็นได้อย่างชัดเจนว่าชั่วโมงของในแต่ละวันมีความสำคัญต่อการเกิดความล่าช้าของเที่ยวบิน



รูปที่ 3 แสดงจำนวนเที่ยวบินที่เกิดความล่าช้าในแต่ละชั่วโมง

6.2 การแปลงคุณสมบัติของข้อมูล (Feature Transformation)

การสร้างแอตทริบิวต์ใหม่โดยใช้แอตทริบิวต์ที่มีอยู่ซึ่งแอตทริบิวต์ใหม่เหล่านี้อาจมีการตีความที่แตกต่างจากแอตทริบิวต์ดั้งเดิม

6.2.1 การสร้างแอตทริบิวต์ใหม่(Create New Feature)

แอตทริบิวต์ hour นำข้อมูลจาก แอตทริบิวต์ SCHEDULED_DEPARTURE มาแบ่งเป็นช่วงในแต่ละช่วงโมง
แอตทริบิวต์ hour_class นำข้อมูลจาก แอตทริบิวต์ใหม่ที่สร้างชื่อ hour ขึ้นมาแบ่งเป็นคลาสโดยอิงวิธีการแบ่งจากกราฟที่แสดงในขั้นตอน Data visualization ซึ่งจะแบ่งออกเป็น 4 คลาส

ระดับ	ชั่วโมง
ต่ำที่สุด (H_lowest)	22,5,23,0,1,2,3,4
ต่ำ (H_low)	8, 20,9,7,6,21
ปานกลาง (H_medium)	14, 12, 11, 10
สูง (H_high)	17,15,19,18,16,13

แอตทริบิวต์ day_delay นำข้อมูลจาก แอตทริบิวต์ DAY_OF_WEEK ขึ้นมาแบ่งเป็นคลาสโดยอิงวิธีการแบ่งจากกราฟที่แสดงในขั้นตอน Data visualization ซึ่งจะแบ่งออกเป็น 3 คลาส

ระดับ	วันที่ในหนึ่งสัปดาห์
ต่ำ (d_low)	6
ปานกลาง (d_medium)	1,2,3,7
สูง (d_high)	4,5

แอตทริบิวต์ month_class นำข้อมูลจาก แอตทริบิวต์ MONTH ขึ้นมาแบ่งเป็นคลาสโดยอิงวิธีการแบ่งจากกราฟที่แสดงในขั้นตอน Data visualization ซึ่งจะแบ่งออกเป็น 3 คลาส

ระดับ	เดือน
ต่ำ (M_low)	9,10,11
ปานกลาง (M_medium)	2,4,5
สูง (M_high)	1,3,6,7,8,12

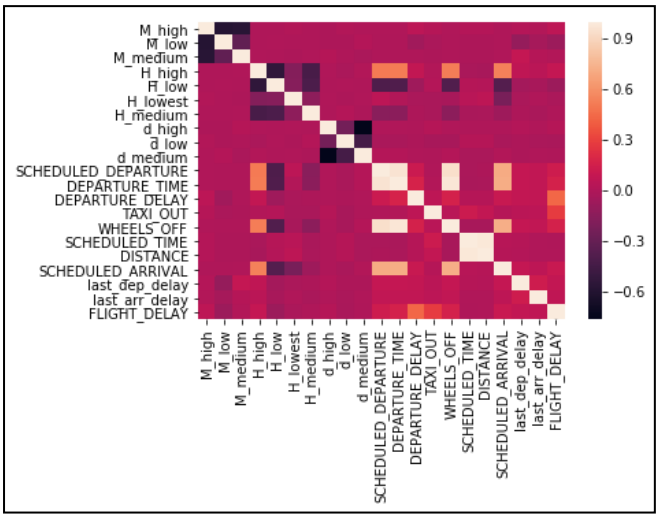
แอตทริบิวต์ last_dep_delay สร้างแอตทริบิวต์ใหม่เพื่อบอกว่าภายในสนามบินต้นทางเดียวกันในระยะเวลา 30 นาทีก่อนหน้ามีเที่ยวบินที่ออกเดินทางล่าช้ากว่ากำหนดกี่นาทีเมื่อเทียบกับเที่ยวบินปัจจุบัน

แอตทริบิวต์ last_arr_delay สร้างแอตทริบิวต์ใหม่เพื่อบอกว่าภายในระยะเวลา 30 นาทีก่อนหน้ามีเที่ยวบินที่มีปลายทางเป็นสนามบินเดียวกันกับสนามบินต้นทางของเที่ยวบินปัจจุบันและมีความล่าช้าเกินกำหนดกี่นาทีเมื่อเทียบกับเที่ยวบินปัจจุบัน

6.2.3 การทำความสะอาดข้อมูล (Data cleansing) เนื่องข้อมูลที่ต้องการ คือ ข้อมูลเที่ยวบินที่มีความล่าช้ากับข้อมูลเที่ยวบินที่บินตามกำหนดแผนการจากมีข้อมูลที่เป็นเที่ยวบินที่ถูกยกเลิก ซึ่งไม่ใช่ข้อมูลที่อยู่ในกลุ่มที่สนใจจึงนำออกและลบข้อมูลที่เป็นค่า missing value ออกจากชุดข้อมูล

6.2.2 one-hot encoding การทำค่าของแอตทริบิวต์ที่เป็นหมวดหมู่ให้อยู่ในรูปแบบที่เป็นตัวเลขเพื่อให้สามารถนำแอตทริบิวต์นั้นเข้าไปเป็นข้อมูลในชุดฝึกฝนได้ ซึ่งในงานวิจัยนี้จะมี 3 แอตทริบิวต์ ที่ถูกนำมาใช้เทคนิค one-hot encoding คือ month_class, hour_class, day_delay

6.2.3 ผลการประเมินความสัมพันธ์ของแอตทริบิวต์ โดย correlation heat map ก่อนที่จะนำข้อมูลเข้าไปใช้ในการฝึกฝน จะพบว่าแอตทริบิวต์ที่มีค่าความสัมพันธ์ที่ทำให้เกิดความล่าช้ามากที่สุดคือ DEPARTURE_DELAY และ TAXI_OUT รองลงมาตามลำดับ



รูปที่ 4 แสดงความสัมพันธ์ของข้อมูลด้วย Correction Heat map

6.3 การสร้างโมเดล (Modeling)

ในขั้นตอนนี้จะทำการสร้างโมเดลโดยใช้เทคนิค Logistic Regression, Gradients Boosting, Ensemble(GB, LG, SVM), CATBoost, Random Forest, Support Vector Machine เพื่อการทำนาย ซึ่งการกำหนดข้อมูลที่ใช้ในการทำนายโอกาสความล่าช้าของเที่ยวบินโดย สุ่มข้อมูลมา 0.01 % (58,191 แถว) จากข้อมูลทั้งหมด (5,819,079 แถว) ในขั้นตอนการเตรียมข้อมูลให้มีความพร้อมสำหรับใช้ในการสร้างโมเดล การแบ่งชุดข้อมูลในการทำนายจะเป็น 80% ที่เป็นชุดข้อมูลสำหรับฝึกฝน และ 20% เป็นข้อมูลสำหรับการทดสอบ ซึ่งโมเดลที่ได้จะทำการประมวลผลเพื่อทำนายให้ได้ค่าความถูกต้อง

ตารางที่ 3 รายละเอียดข้อมูลหลังจากทำการแปลงคุณสมบัติข้อมูล

	ตัวอย่าง	แอตทริบิวต์	คลาส 0	คลาส 1
ข้อมูลทั้งหมด	56,375	20	35,873	20,502

7. ผลการดำเนินงานวิจัย

งานวิจัยฉบับนี้ได้จัดทำเทคนิคต่างๆเพื่อสร้างโมเดลการทำนายความล่าช้าของเที่ยวบินโดยสร้างโมเดลจากเทคนิคทั้ง 6 เทคนิคดังนี้ Logistic Regression, Gradients Boosting, Ensemble(GB, LG, SVM), CATBoost, Random Forest, Support Vector Machine

ตารางที่ 4 แสดงผลการเปรียบเทียบค่าความถูกต้อง(Accurency)

โมเดล	GB	EN	LG	CB	RF	SVM
ค่าความถูกต้อง (Accuracy)	84.2%	86.3%	86.6%	86.8%	85.5	86.0%

ตารางที่ 5 รายละเอียดโมเดลที่ทดลอง

รายละเอียดโมเดล	ชื่อย่อ
Logistic Regression	GB
Ensemble(GB, LG, SVM)	EN
Gradients Boosting	LG
CATBoos	CB
Random Forest	RF
Support vector machine	SVM

จากตารางที่ 4 เป็นการแสดงการเปรียบเทียบค่าความถูกต้องที่ได้จากการสร้างโมเดลในการทำนายโอกาสที่จะเกิดความล่าช้าของเที่ยวบินด้วยเทคนิคทั้ง 6 เทคนิค ผลที่ได้จากการสร้างโมเดลเพื่อการทำนายโดยแบ่งข้อมูลเป็นชุดข้อมูลสำหรับการทดสอบ 20 % ของการทดสอบทั้งหมด และชุดข้อมูลสำหรับฝึกฝน 80 % ของข้อมูลทั้งหมด [80:20] โดยเทคนิคที่ให้ค่าความถูกต้องมากที่สุดคือ เทคนิค CAT Boost คิดเป็นร้อยละ 86.6 %

8. สรุปผลการทดลอง

งานวิจัยนี้ เป็นการสร้างโมเดลเพื่อการทำนายโอกาสที่จะเกิดความล่าช้าของเที่ยวบิน โดยใช้ข้อมูลตัวอย่างจาก ข้อมูลเที่ยวบิน องประเทศสหรัฐอเมริกา ในปี 2015 โดยมีจำนวนข้อมูลทั้งหมด 5,819,079 แถว โดยใช้วิธีการและเทคนิคต่างๆ เพื่อคัดเลือกข้อมูลที่เป็นปัจจัยส่งผลให้เกิดความล้าช้า จนสามารถนำข้อมูลไปสร้างเป็นองค์ความรู้ที่ต้องการ ซึ่งนำมาสร้างการทำนายหาโอกาสการเกิดความล่าช้าของเที่ยวบิน โดยใช้ข้อมูลที่ส่งผลจำนวน 20 แอตทริบิวต์ ได้แก่ เวลาออกเดินทางตามแผน (SCHEDULED_DEPARTURE) เวลาที่เครื่องออกเดินทางจริง (DEPARTURE_TIME) ความล่าช้าที่เครื่องออกเดินทางจริงกับกำหนดการ (DEPARTURE_DELAY) ระยะเวลาที่ใช้ระหว่างเกตไปกระทั่งเวลาที่ล้อเครื่องถูกเก็บ (TAXI_OUT) ระยะเวลาที่ใช้ตามแผน (SCHEDULED_TIME) เวลาที่ล้อถูกเก็บขึ้น (WHEELS_OFF) เวลาถึงที่หมายตามแผน (SCHEDULED_ARRIVAL) ระยะทางการเดินทาง (DISTANCE) คลาสของช้าวโมงที่มีความล่าช้า (hour_class: : ต่ำ (h_low) ปานกลาง (h_medium) สูง (h_high)) คลาสของวันที่มีความล่าช้า (day_delay: ต่ำ (d_low) ปานกลาง (d_medium) สูง (d_high)) คลาสของเดือนที่มีความล่าช้า (month_class: ต่ำ (m_low) ปานกลาง (m_medium) สูง (m_high) (ความล่าช้าที่เครื่องออกเดินทางจริงกับกำหนดการของเที่ยวบินก่อนหน้า) last_dep_delay (ความล่าช้าของเที่ยวบินก่อนหน้าที่มีท่าอากาศยานปลายทางเป็นท่าอากาศยานเดียวกับท่าอากาศยานต้นทางของเที่ยวบินปัจจุบัน)last_arr_delay ซึ่งจะนำข้อมูลเหล่ามาใช้กับโมเดลการทำนายที่ใช้เทคนิค Logistic Regression, Gradients Boosting, Ensemble(GB, LG, SVM), CATBoost, Random Forest และ Support Vector Machine เพื่อเปรียบเทียบค่าความถูกต้องของโมเดลแต่ละประเภท โดยมีการแบ่งชุดข้อมูล 80% เป็นชุดข้อมูลฝึกฝน และ 20% เป็นชุดทดสอบ ซึ่งการสร้างโมเดลที่ใช้เทคนิคที่มีประสิทธิภาพในการประเมินค่าความถูกต้องได้มากที่สุดคือ CATboost คิดเป็นร้อยละ 86.8%

จากผลงานวิจัยคาดว่าสามารถนำโมเดลไปใช้เพื่อช่วยในการตัดสินใจในการจัดการเที่ยวบินเพื่อลดปัญหาที่ทำให้เกิดความล่าช้าสะสมได้ แต่ยังไม่สามารถนำไปใช้ตัดสินใจแทนได้ทั้งหมดเนื่องจากค่าความถูกต้องสูงสุดได้เพียง86.8% ซึ่งควรได้รับค่าความถูกต้องอย่างน้อย95%ถึงจะอยู่ในจุดที่ยอมรับได้ เพราะว่ากำหนดเที่ยวบินนั้นมีผลกระทบอย่างมากต่อสายการบินและท่าอากาศยาน รวมถึงผู้โดยสารอย่างมาก การนำไปใช้จริงจึงต้องการความถูกต้องที่สูงและต้องการแอตทริบิวต์ที่ใช้ในการวิจัยที่มีความสำคัญเพิ่มขึ้นเพื่อเพิ่มค่าความถูกต้อง เช่น ข้อมูลสภาพอากาศ ข้อมูลจำนวนผู้โดยสารในเที่ยวบิน ที่อาจมีความสัมพันธ์กับความล่าช้าของเที่ยวบิน

เอกสารอ้างอิง

[1] Roshni Musaddi 1, Anny Jaiswal 2, Pooja J 3, Mansvi Girdonia 4, Minu M.S, “Flight Delay Prediction using Binary Classification,” International Journal of Emerging Technologies in Engineering Research (IJETER) Volume 6, Issue 10, October (2018)

[2] ยุทธ ไกยวรรณ์, “หลักการและการใช้การวิเคราะห์การถดถอยโลจิสติคสำหรับการวิจัย,” วารสารวิจัยมหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย 4(1) : 1-12 (2555)

[3] H. Khaksar and A. Sheikholeslami, “Airline delay prediction by machine learning algorithms,” Scientia Iranica A (2019)