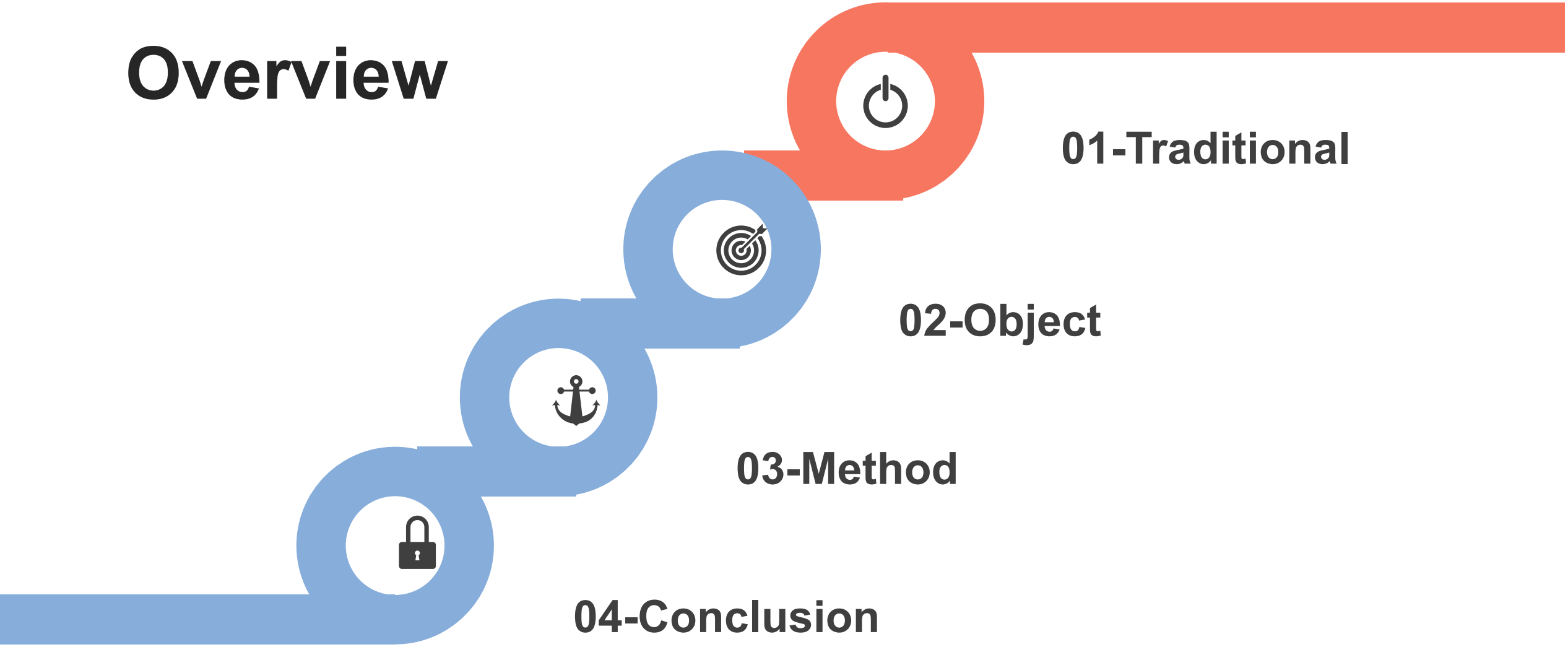
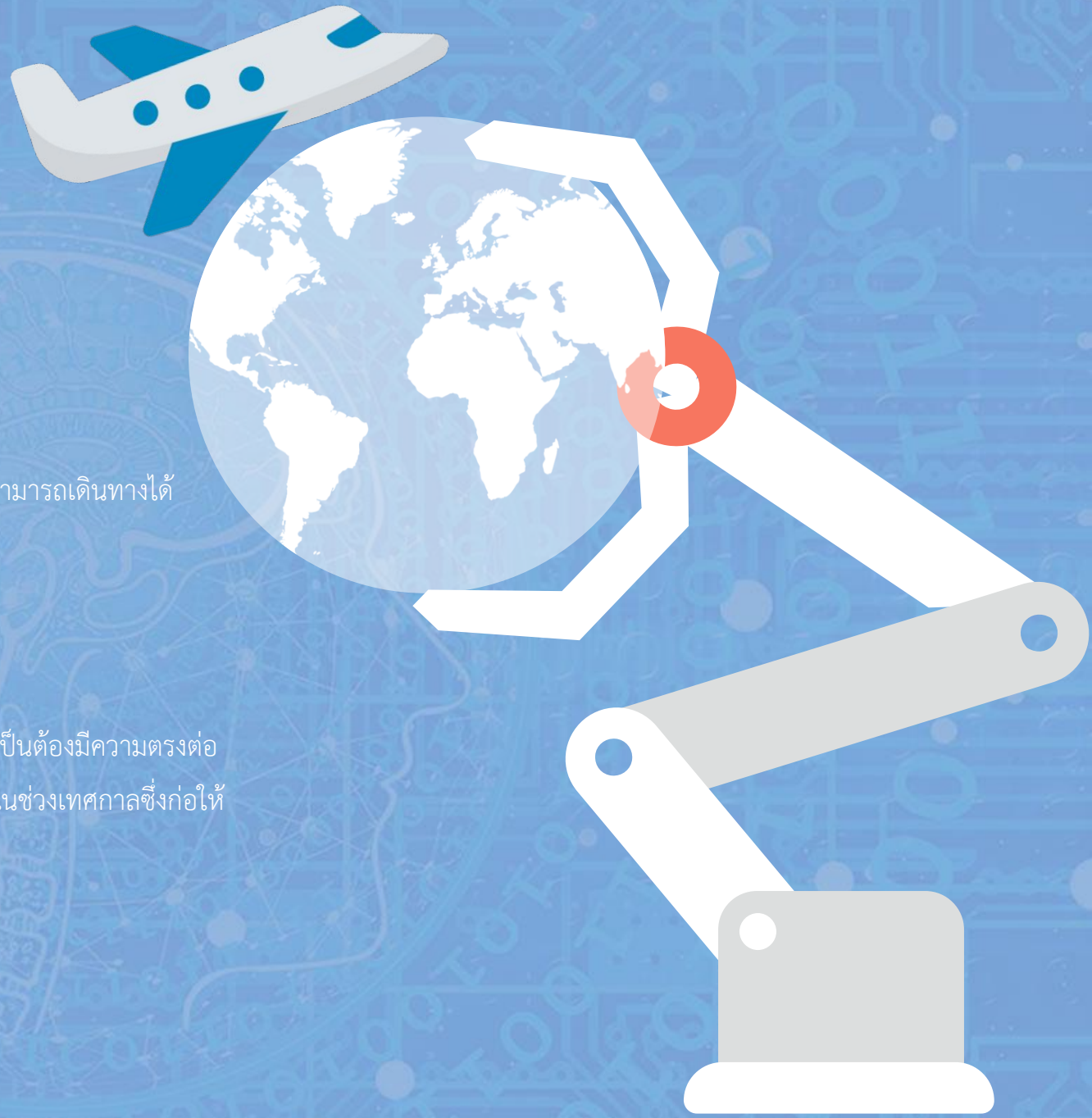
Building the Predicting model of Flight Delay Using Boosting Algorithm

By
Panida Katklangdon

Overview



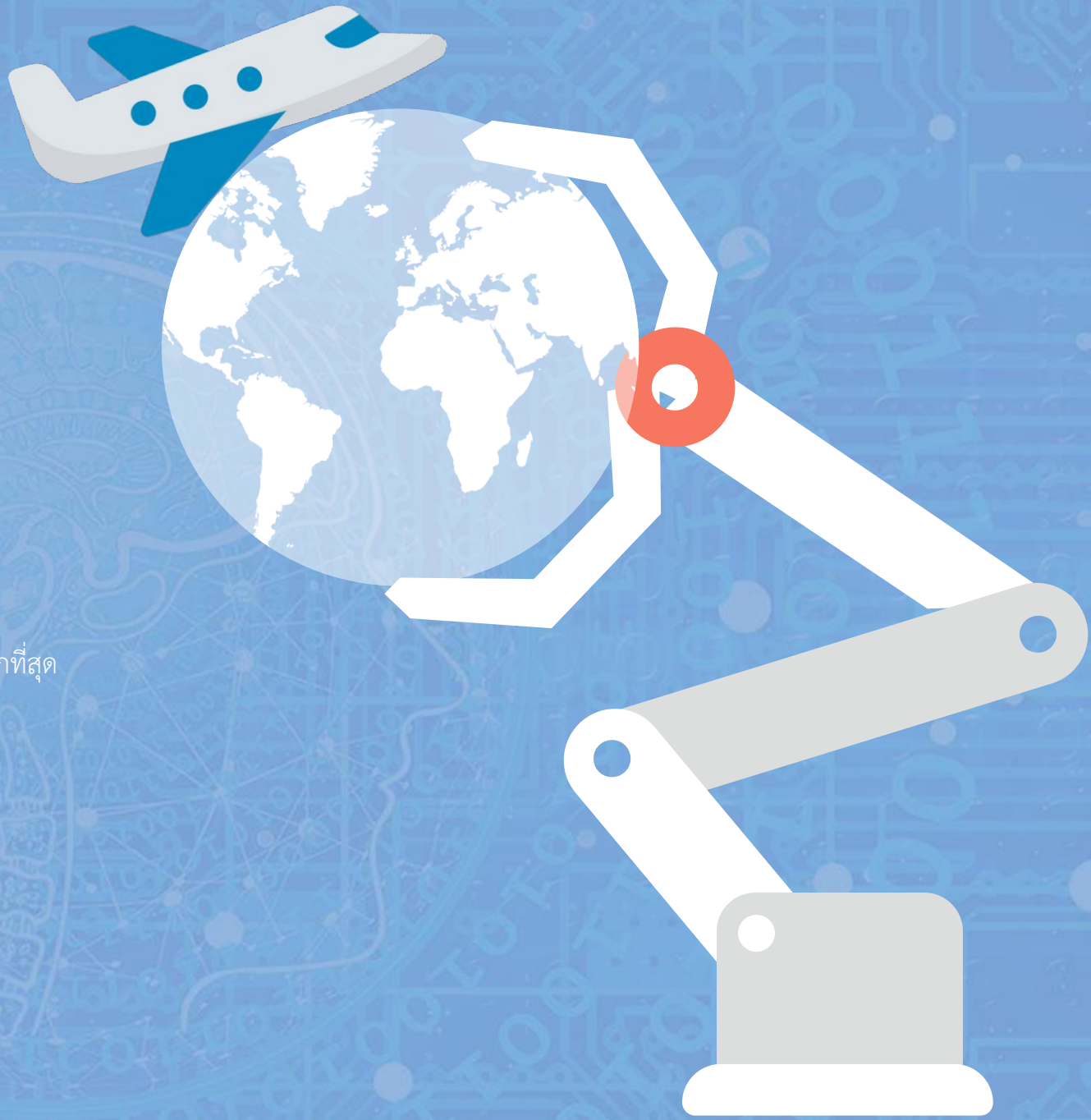
Traditional



- เนื่องจากปัจจุบันการเดินทางระยะไกลส่วนใหญ่เป็นการเดินทางโดยเครื่องบินเพราะว่าสามารถเดินทางได้อย่างรวดเร็วเมื่อเทียบกับการเดินทางของทางบกและทางน้ำ

- ดังนั้นผู้โดยสารนั้นให้ความสำคัญของเวลาเป็นอย่างมากซึ่งผู้ให้บริการสายการบินนั้นจำเป็นต้องมีความตรงต่อเวลา ปัจจุบันเครื่องบินดีเลย์นับเป็นปัญหาใหญ่ของการขนส่งทางอากาศ โดยเฉพาะในช่วงเทศกาลซึ่งก่อให้เกิดปัญหาความไม่สะดวกต่อผู้โดยสารเป็นอย่างมาก

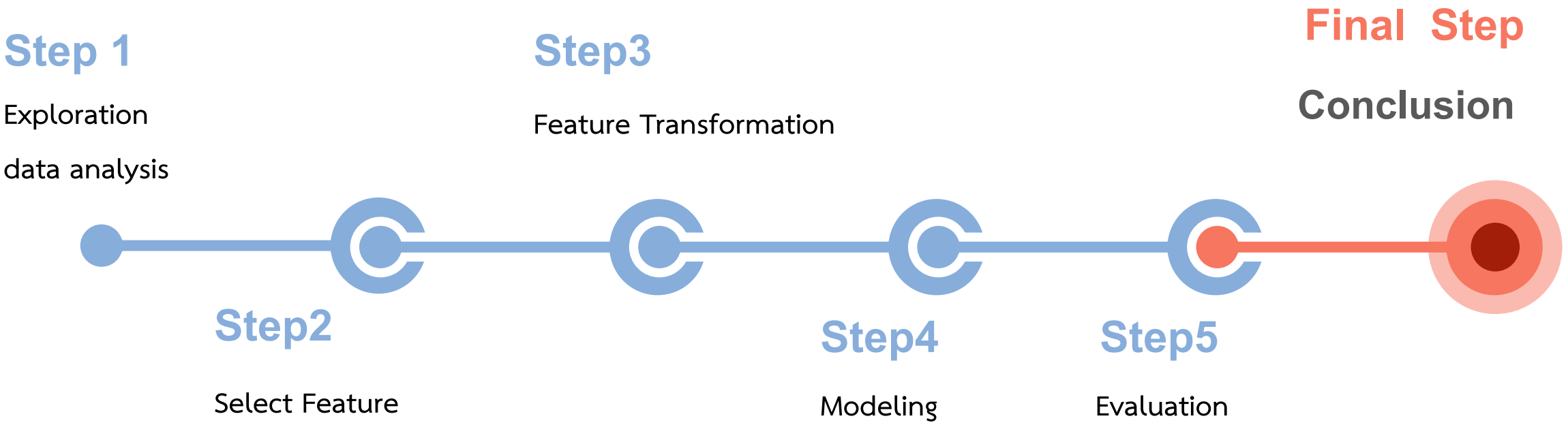
Object



- เพื่อสร้างโมเดลการทำนายโดยคัดเลือกเทคนิคที่ให้ประสิทธิภาพของความแม่นยำมากที่สุด
- เพื่อทำนายหาโอกาสการเกิดความล่าช้าในแต่ละเที่ยวของเที่ยวบิน

- เพื่อทำนายหาโอกาสการเกิดความล่าช้าในแต่ละเที่ยวของเที่ยวบิน

Method


Step 1
Exploration
data analysis
Step2
Select Feature
Step3
Feature Transformation
Step4
Modeling
Step5
Evaluation
Final Step
Conclusion

Exploration

data analysis

รายละเอียดข้อมูล ข้อมูลแต่ละแถวจะเป็นข้อมูล
ตั้งแต่เครื่องออกจากสนามบินต้นทางไปจนถึงสนามบินปลายทาง

จำนวนข้อมูลทั้ง
5,819,079 แถว 31 คอลัมน์

- ทำการสุ่มตัวอย่างข้อมูลมา 1%(5หมื่นแถว) จากข้อมูลทั้งหมด 5ล้านแถว เนื่องจาก kernel มีปัญหาล้มเหลวบ่อยเนื่องจากฝึกฝนโมเดลกับข้อมูลจำนวนมาก
- ได้ทำการทดสอบข้อมูลทั้งหมด6ขนาดแล้วคือ 1%, 5%, 10%, 20%, 50%, 100%, พบว่าไม่มีผลกับการสร้างโมเดลเนื่องจากข้อมูลคือเที่ยวบินที่เริ่มตั้งแต่วันที่ 1 ม.ค. 2015 ถึง 31 ธ.ค. 2015 ซึ่งเป็นข้อมูลที่มีการเรียงลำดับมาเรียบร้อยแล้ว ทำให้ข้อมูลที่สุ่มตัวอย่างในแต่ละขนาดจึงมีจำนวนในแต่ละเดือนเท่าๆกัน

- ข้อมูลที่ใช้จะเป็นข้อมูลที่สามารถได้รับ

ตั้งแต่เครื่องบินเริ่มเก็บล้อและลอยตัวออกจากรันเวย์

ข้อมูลต่างๆที่ได้รับหลังจากเก็บล้อ จะทำการลบออกจากชุดข้อมูล

Exploration

data analysis(2)

['CANCELLATION_REASON', 'CANCELLED',

'ARRIVAL_TIME', 'DIVERTED', 'ELAPSED_TIME',

'AIR_TIME', 'WHEELS_ON', 'TAXI_IN', 'AIR_SYSTEM_DELAY'

, 'SECURITY_DELAY', 'AIRLINE_DELAY', 'LATE_AIRCRAFT_DELAY', 'WEATHER_DELAY']

ชุดคอลัมน์ข้างบนคือข้อมูลที่จะได้รับตั้งแต่เครื่องเริ่มลอยตัวจากรันเวย์จนถึงท่าอากาศยานปลายทางซึ่งเป็นข้อมูลที่ไม่สามารถใช้ได้ในการสร้างโมเดลได้ เพราะจุดมุ่งหมายของโปรเจคนี้คือการทำนายตอนที่เครื่องบินเก็บล้อเท่านั้น

['ORIGIN_AIRPORT' 'DESTINATION_AIRPORT' 'TAIL_NUMBER' 'FLIGHT_NUMBER'] คือชุดคอลัมน์ข้อมูลที่จะเป็นแบบคลาสซึ่งในแต่ละคอลัมน์มีประเภทข้อมูลที่หลากหลายและไม่สามารถนำไปทำ one-hot เพื่อใช้เป็นข้อมูลในการสร้างโมเดลได้

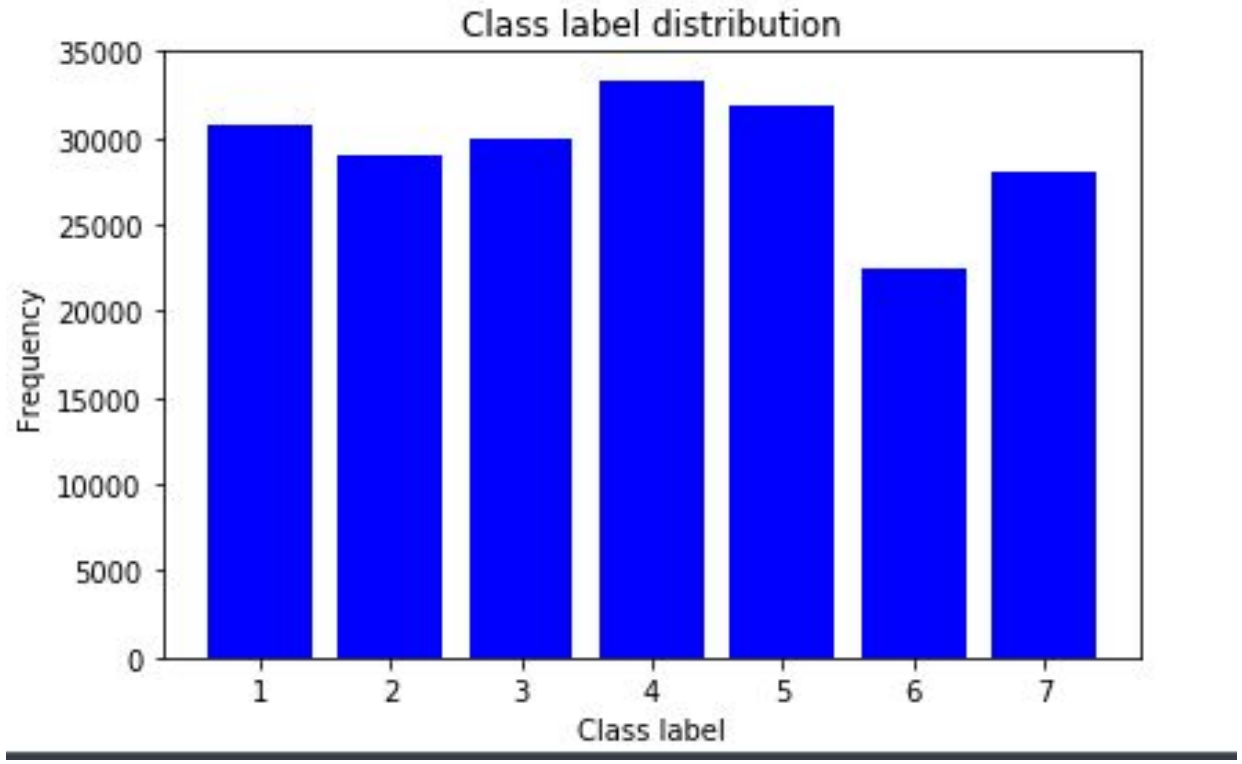
Select Feature

- **จำนวนแอตทริบิวต์ที่ถูกเลือกมาทั้งหมด 12 แอตทริต์**

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

Data visualization & Feature Transformation

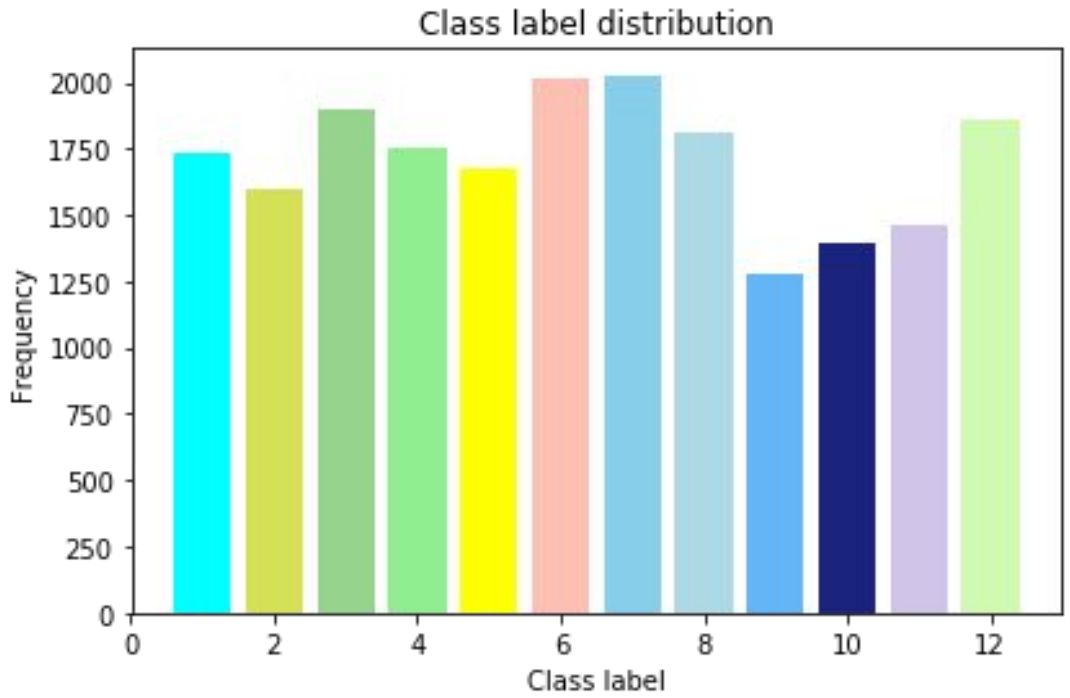
แสดงจำนวนเที่ยวบินที่เกิดความล่าช้าในแต่ละวันของสัปดาห์



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

Data visualization & Feature Transformation

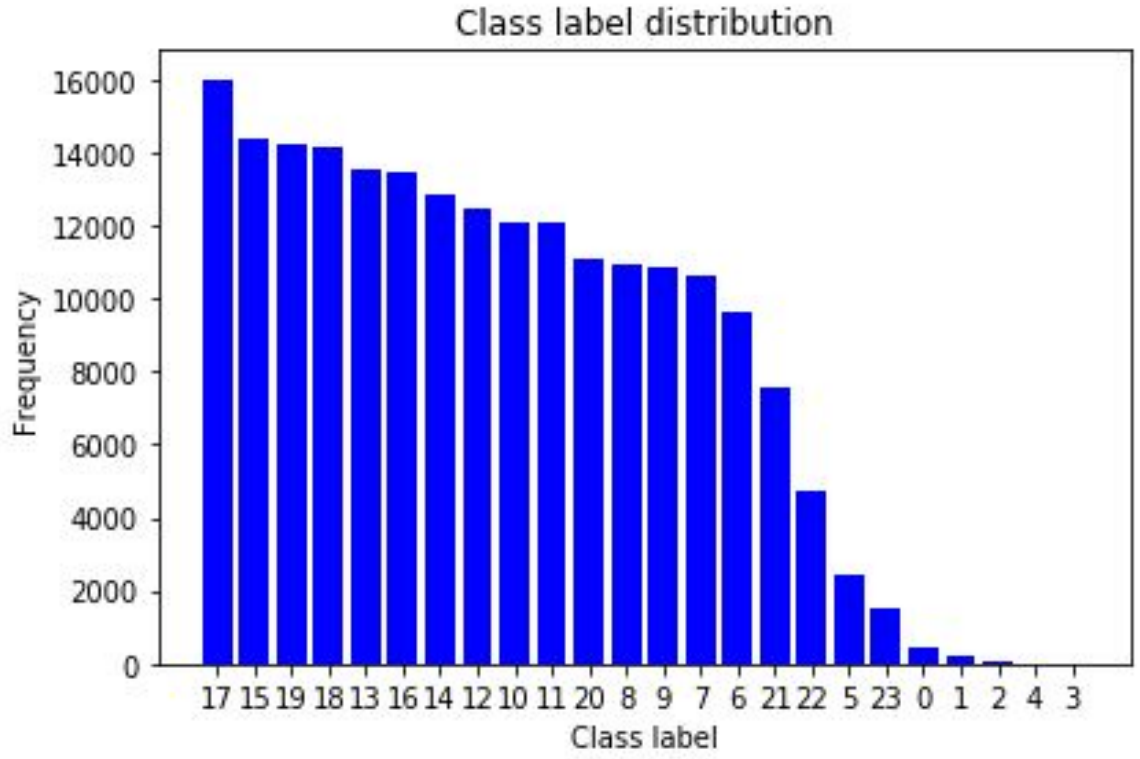
แสดงจำนวนเที่ยวบินที่เกิดความล่าช้าในแต่ละเดือน



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

Data visualization & Feature Transformation

แสดงจำนวนเที่ยวบินที่เกิดความล่าช้าในแต่ละชั่วโมง



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

Data visualization & Feature Transformation

one-hot encoding

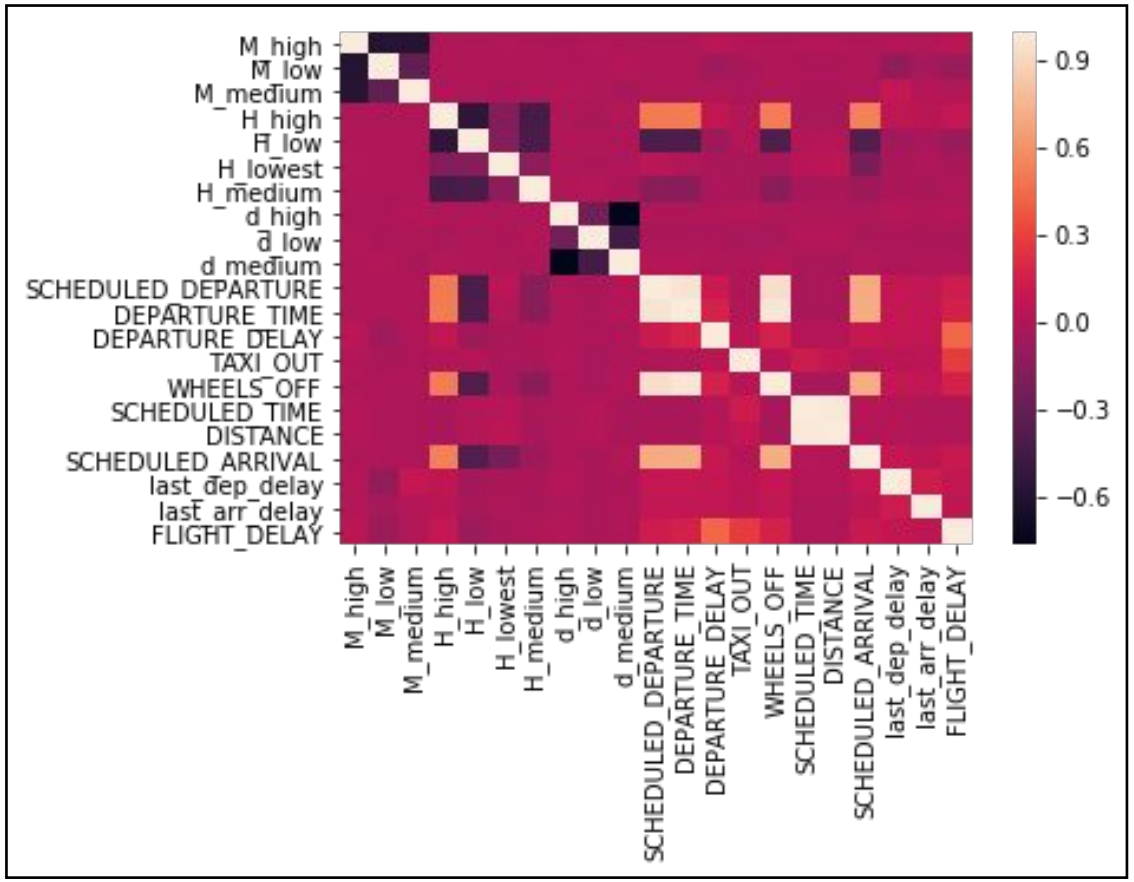
3 แอตทริบิวต์ ที่ถูกนำมาใช้เทคนิค one-hot encoding คือ

- month_class
- hour_class
- day_delay

ผลการประเมินความสัมพันธ์ของแอตทริบิวต์

- แอตทริบิวต์ที่มีค่าความสัมพันธ์ที่ทำให้เกิดความล่าช้ามากที่สุดคือ DEPARTURE_DELAY และ TAXI_OUT รองลงมาตามลำดับ

Correlation heatmap



Modeling(การสร้างโมเดล)

ชุดข้อมูลทั้งหมดที่ใช้ในการฝึกฝนและทดสอบโมเดล

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

แบ่งข้อมูลทั้งหมดเป็น 3 ชุด

train(64%) valid(16%) test (20%)

สร้างโมเดลโดยใช้เทคนิค

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

ผลการดำเนินงานวิจัย

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

การสร้างโมเดลที่ใช้เทคนิคที่มีประสิทธิภาพในการประเมินค่าความถูกต้องได้มากที่สุดคือ CATboost คิดเป็นร้อยละ 86.8%

สรุปผลการทดลอง

จากผลงานวิจัยคาดว่าสามารถนำโมเดลไปใช้เพื่อช่วยในการตัดสินใจในการจัดการเที่ยวบินเพื่อลดปัญหาที่ทำให้เกิดความล่าช้าสะสมได้
แต่ยังไม่สามารถนำไปใช้ตัดสินใจแทนได้ทั้งหมดเนื่องจากค่าความถูกต้องสูงสุดได้เพียง86.8%
ซึ่งควรได้รับค่าความถูกต้องอย่างน้อย95%ถึงจะอยู่ในจุดที่ยอมรับได้
เพราะว่ากำหนดเที่ยวบินนั้นมีผลกระทบอย่างมากต่อสายการบินและท่าอากาศยาน
รวมถึงผู้โดยสารอย่างมาก การนำไปใช้จริงจึงต้องการความถูกต้องที่สูงและต้องการแอตทริบิวต์ที่ใช้ในการวิจัยที่มีความสำคัญเพิ่มขึ้น
เพื่อเพิ่มค่าความถูกต้อง
เช่นข้อมูลสภาพอากาศ ข้อมูลจำนวนผู้โดยสารในเที่ยวบิน ที่อาจมีความสัมพันธ์กับความล่าช้าของเที่ยวบิน