สรุปเนื้อหาการนำเสนอ รายวิชา 05177401 : สัมมนาทางด้านวิทยาการข้อมูล 1

Translated & Interpreted by Puriwat Sangrawee (นายภูริ วัฒน์ แสงระวี) / Student ID 65056071

Research Article : Lead Time Prediction Using ML Algorithms : A case Study by a semiconductor Manufacturer

Author : Lukas Lingitz et. al. (2018) via ScienceDirect

บทคัดย่อ (Abstract)

การทำนายอย่างแม่นยำสำหรับ Manufacturing Lead Time ส่งผลอย่างมีนัยสำคัญต่อคุณภาพ (Quality) และประสิทธิภาพ (Efficiency) ของการวางแผนและจัดตารางการผลิต (Production Planning & Scheduling : PPS) ซึ่งสำหรับวิธีการวางแผนและควบคุมการผลิตแบบดั้งเดิมนั้น โดยมาก ค่า Lead Time เฉลี่ย จะได้มาจากข้อมูลการผลิตย้อนหลัง (Historical Data) ซึ่งมักจะนำไปสู่ความบกพร่องของระบบ PPS เนื่องมาจากฝ่ายวางแผนการผลิตมิได้คำนึงถึงเรื่องความแปรปรวน (Variation) ของ Lead Time ซึ่งเป็นผลมาจากหลากหลายข้อกำหนดในแวดวงอุตสาหกรรมการผลิตในยุคปัจจุบัน

ซึ่งสำหรับในเคสของการผลิตเซมิคอนดักเตอร์นี้ วิธีการทำนาย Lead Time ที่ซับซ้อนและครอบคลุมถือเป็นสิ่งจำเป็น อันเนื่องจากจากในกระบวนการผลิตเซมิคอนดักเตอร์ประกอบด้วย กระบวนการการผลิตที่ซับซ้อน , มีการผลิตจำนวนมาก ,มีหน่วยการผลิตย่อยต่อไลน์การผลิตหลายหน่วย รวมทั้งยังจำเป็นต้องคงไว้ซึ่งประสิทธิภาพระดับสูงในกระบวนการทำงาน เพื่อที่จะเอาชนะทุกความท้ายทายเหล่านี้จึงมีการหยิบยกเอา Supervised ML เข้ามาประยุกต์ใช้ในการทำนาย Lead Time โดยอาศัยข้อมูล Historical Production Data จากระบบการดำเนินการผลิต (manufacturing Execution System : MES) โดยในงานวิจัยฉบับนี้ ทีมผู้วิจัยได้ทำการตรวจสอบการใช้ Regression Algorithm ใหม่ล่าสุดที่มีการใช้งานเพื่อที่จะดูว่าการใช้ Algorithm ดังกล่าวนี้ส่งผลต่อความแม่นยำของการทำนาย Lead Time อย่างไรบ้าง ผ่าน use case จริงจากภาคอุตสาหกรรม อีกทั้งการศึกษานี้ยังได้มีการเปรียบเทียบผลจากหลากหลายรูปแบบ และมีการสรุปการเลือกใช้ฟีเจอร์ และการประยุกต์ใช้วิธีการที่เหมาะสมที่สุดกับอุตสาหกรรมเซมิคอนดักเตอร์

1. บทนำ

การวิเคราะห์ข้อมูลเชิงทำนาย (Predictive Data Analytics) นั้น หมายรวมถึงการสร้างและใช้งาน model ที่ทำการทำนาย (Prediction) โดยอาศัยรูปแบบ (Pattern) ที่ได้จาก Historical Data ซึ่งเมื่อพิจารณาถึง 6 key phases ของ Predictive Data Analytics Project Lifecycle ที่ถูกนิยามโดย Cross Industry Standard Process of Data Mining (CRISP-DM) อันประกอบไปด้วย (1) Business Understanding (ความเข้าใจในธุรกิจ) (2) Data Understanding (ความเข้าใจในข้อมูลที่ศึกษา) (3) Data Preparation (การจัดเตรียมข้อมูล) (4) Modeling (การสร้างโมเดลสำหรับการทำนาย) (5) Evaluation (การประเมินประสิทธิภาพโมเดล) และ (6) Deployment (การนำโมเดลไปใช้งานจริง)

ซึ่งหากกล่าวถึงในส่วนของ phase ที่ 4 (Modeling) นั้นก็คือส่วนที่ ML ถูกนำมาใช้งานเพื่อสร้าง Predictive Model โดยโมเดลที่ดีที่สุดจะถูกประเมินประสิทธิภาพ เพื่อพิสูจน์ว่าเหมาะสมผ่านการนำไปใช้งานจริง เช่น ในกรณีนี้จะนำไปใช้ร่วมกับระบบ MES หรืออาจกล่าวได้ว่า ML ถูกนิยามในฐานะกระบวนการอัตโนมัติ (Automated Process) ซึ่งทำการหารูปแบบ (Pattern) จาก historical data โดย ณ ขณะนี้ ใครซักคนอาจจะแยกความแตกต่างระหว่าง 2 รูปแบบหลักได้แล้ว นั่นคือ

(1) **การเรียนรู้แบบมีผู้สอน (Supervised ML)** ซึ่งถูกตั้งสมมติฐานว่าชุดข้อมูลฝึกหัด (Training Example) นั้นถูกจำแนกประเภท (Classified หรืออาจใช้คำว่า labelled) (อาทิ การศึกษาความสัมพันธ์ ระหว่าง set ของ Descriptive feature และ target feature)

(2) **การเรียนรู้แบบไม่มีผู้สอน (Unsupervised ML)** เกี่ยวข้องกับการวิเคราะห์ตัวอย่างที่มิได้จำแนกประเภท (Unclassified Samples)

นอกจาก 2 ประเภทหลักดังกล่าวแล้ว ML ยังมีรูปแบบอื่น ๆ อีกหลากหลาย ไม่ว่าจะเป็น การเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning) หรือการเรียนรู้แบบเสริมกำลัง (Reinforcement Learning) แต่สำหรับในงานวิจัยชุดนี้ เราจะพิจารณาเฉพาะการเรียนรู้แบบมีผู้สอน (Supervised ML) โดยเจาะจงเฉพาะขั้นตอนวิธีการถดถอย (Regression Algorithms)

2. การทำนายค่า Lead Time ด้วย ML Algorithms

2.1 การค้นพบความรู้ในการวางแผนและควบคุมการผลิต

สถิติ , การทำเหมืองข้อมูล หรือการค้นพบความรู้ ถูกนิยามครั้งแรกในปี 1989 ในฐานะเครื่องมืออัจฉริยะรูปแบบใหม่ในการสกัดเอาความข้อมูลและความรู้ที่มีประโยชน์ออกมา (ซึ่งจะเรียกว่า Actionable information หรือ Hidden Pattern) จากฐานข้อมูลที่แตกต่างกัน โดยในครั้งแรกนั้น การค้นพบความรู้ถูกนำไปประยุกต์ใช้อย่างกว้างขวาง ในหลากหลายวงการ ไม่ว่าจะเป็น ด้านการแพทย์ , การเงิน ,เทคโนโลยีชีวภาพ หรือการตลาด แต่กระนั้นการศึกษาวิจัยด้านนี้ที่เกี่ยวข้องกับอุตสาหกรรมการผลิตยังถือว่าน้อย แต่อย่างไรก็ดี ใน 2-3 ปีมานี้ได้มีการเติบโตขึ้นของจำนวนงานวิจัยที่อภิปรายถึงขั้นตอนวิธี รวมถึงเทคนิคในการวิเคราะห์ข้อมูลในการจัดการการผลิตอย่างมีนัยสำคัญ และจากการอ้างอิงจากงานวิจัยของ ***Rainer*** ภายหลังการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลที่แตกต่างกันสำหรับการเปลี่ยน Big Data ให้กลายมาเป็น Smart Data แล้ว บริษัทนั้นสามารถได้ทุนคืนเป็นอย่างน้อย 10 เท่าจากเงินลงทุนที่ใช้

สำหรับ output data ของกระบวนการทำเหมืองข้อมูลนั้นสามารแบ่งได้เป็น 2 ประเภทหลัก ตามหน้าที่การทำงาน และตามเป้าหมายของเทคนิคที่ใช้ โดย

การวิเคราะห์สถิติเชิงพรรณนา (Descriptive Statistical Analytics) (อย่างเช่น การวิเคราะห์ความสัมพันธ์ (Association Analysis) หรือการจัดกลุ่ม (Clustering)) จะเน้นไปที่การค้นคว้ากฎหรือรูปแบบเพื่อที่จะอธิบายข้อมูล ในขณะที่

การทำเหมืองข้อมูลเชิงทำนาย (Predictive Data Mining) (อย่างเช่น การจำแนก (Classification) หรือ การถดถอย (Regression)) จะใช้สำหรับวิเคราะห์สิ่งที่เกี่ยวข้องและข้อมูลจริง (Actual data) เพื่อที่จะทำนายค่าของตัวแปรเป้าหมาย (Key variable) (ซึ่งอาจมีเพียงตัวแปรเดียวหรือมากกว่า) ที่เป็นไปได้ในอนาคต โดยที่ Regression ถือเป็นรูปแบบหนึ่งของกระบวนการ Predictive Data Mining ซึ่งมีวัตถุประสงค์เพื่อทำนายค่าของตัวแปรต่อเนื่อง (Continuous variable)

เฉิง และคณะ (Cheng et al.) ได้ทำการแก้ไขเพิ่มเติมงานวิจัยที่เกี่ยวข้องในปี 2010 และได้อภิปรายถึงเทคนิคเบื้องต้นในการทำเหมืองความรู้ในกระบวนการจัดการผลิต ตามการสำรวจดังกล่าวนี้พบว่า ขอบข่ายของงานที่สะท้อนออกมาในรูปของผลลัพธ์ดังกล่าวนี้มากที่สุด 4 อันดับแรก ได้แก่ การวางแผนและจัดตารางการผลิตขั้นสูง (Advanced planning & Scheduling) , การปรับปรุงคุณภาพการผลิต (Quality Improvement) , การวินิจฉัยข้อผิดพลาด (Fault Diagnosis) และการวิเคราะห์ข้อบกพร่องของชิ้นงานในการผลิต (Defect Analysis) สำหรับในรูปแบบที่ 5 นั้น ก็ได้มีการนิยามเกิดขึ้นมาด้วยเช่นกัน ในทำนองว่า flow time/cycle time ที่ทำนายได้จะมีค่าเป็นเท่าใด หากทราบค่า life และประสิทธิภาพการผลิต (yield) ที่ได้จากการทำนาย โดย PPC นั้นถูกระบุในฐานะ research gap ในปี 2009 และถูกนับมาทบทวนอีกรอบในปี 2017 โดยในรอบหลังนี้ได้มีการเปิดเผยการนำไปใช้งานจริงของเทคนิคข้างต้นในบางประการ ในรอบ 9 ปีที่ผ่านมา ซึ่งอาจกล่าวถึงผลลัพธ์ได้ว่า ในแวดวงการวิจัยอาจจะต้องให้ความสนใจมากขึ้นในส่วนของการทำเหมืองข้อมูลสำหรับ PPC

2.2 เทคนิคล่าสุดในการพยกรณ์ Lead time ด้วย Regression

ในการศึกษาปัจจุบันนั้น lead time ในฐานะพารามิเตอร์ควบคุมที่สำคัญที่สุดตัวหนึ่งและในฐานะตัวแปรเป้าหมาย (target figure) ของ PPC นั้นได้ถูกวิเคราะห์และทำนายด้วยความช่วยเหลือของ ML Algorithms หลากหลายรูปแบบที่แตกต่างกัน รวมทั้งข้อมูลจาก MES โดยจากการทบทวนวรรณกรรมนั้นได้เปิดเผยว่า การทำวิจัยส่วนใหญ่ที่เกี่ยวข้องกับเวลานั้นมีความสัมพันธ์กับการวิเคราะห์การทำเหมืองข้อมูล (flow time , (lot) cycle time , lead time) โดยพบการศึกษาหลากหลายรูปแบบด้วยกัน ไม่ว่าจะเป็น i) มุ่งเน้นไปที่ตลอดทั้งกระบวนการไหล (process flow) ii) การใช้ชุดข้อมูล (dataset) ที่ถูกสร้างขึ้นจากการจำลองกระบวนการ (simulation) หรือ iii) มีการประยุกต์ใช้ ML Algorithms และเปรียบเทียบผลลัพธ์ที่ได้จาก Algorithms ต่างชนิดกัน

เฟฟเฟอร์ และคณะ (Pfeiffer et al.) ทำการเปรียบเทียบผลลัพธ์จาก ML ทั้ง 3 model ที่ใช้ในการทำนาย lead time ด้วย 8 ฟีเจอร์จากข้อมูลซึ่งได้มาจากการจำลองเหตุการณ์แบบไม่ต่อเนื่อง (discrete-event simulation) โดยในการศึกษานี้ พวกเขาพบว่า การใช้งานโมเดลแบบ Random forest มีประสิทธิภาพดีกว่าโมเดลแบบ Linear Regression และ Regression Tree

ออสเติร์ก และคณะ (Ozturk et al.) ได้ทำการประยุกต์ใช้ Regression Tree กับแหล่งข้อมูลจำลองของร้านค้า 4 ประเภทเพื่อที่จะหาคุณลักษณะ (attribute) ที่มีความเกี่ยวข้องมากที่สุดซึ่งจะมีความสามารถในการทำนายสูง

เมแดน และคณะ (Meidan et al.) ได้มุ่งเน้นไปที่การใช้ระยะเวลาการคอย (waiting time) แทนที่จะใช้ lead time ทั้งหมด ซึ่งหลังจากการทำให้ทุกตัวแปรแบบต่อเนื่อง (continuous variable) กลายเป็นค่าไม่ต่อเนื่อง (Discrete) โดยการเลือกตัวแปรโดยอาศัยตัวจำแนกแบบนาอีฟเบย์ , ต้นไม้ตัดสินใจ (Decision Tree) , โครงข่ายประสาทเทียม (ANN) และการถดถอยโลจิสติกแบบหลายกลุ่ม (Multinomial Logistic Regression) มาใช้ในการประเมินประสิทธิภาพ จากการประเมินพบว่า 182 ฟีเจอร์จาก Original Dataset ซึ่งเป็นชุดข้อมูลที่ถูกสร้างขึ้นจากการจำลอง (Simulation) นั้นสามารถลดลงไปจนเหลือนำมาใช้งานเพียง 20 ฟีเจอร์ได้

อเลนซี และคณะ (Alenzi et al.) ได้ใช้โมเดล Support Vector Machine Regression Model (SVM) ที่มีการปรับตั้งอย่างละเอียดแล้ว มาทำการทำนายค่าระยะเวลาการไหล (flow time) แบบเรียลไทม์และนำไปเปรียบเทียบผลลัพธ์กับโมเดล ANN และอนุกรมเวลา (Time Series) แบบดั้งเดิม โดยใช้ชุดข้อมูลที่ได้มาจากการจำลอง (Simulation) และพบว่าโมเดลแบบ SVM ให้ผลลัพธ์ดีที่สุด

โมริ และคณะ (Mori et al.) ศึกษาเปรียบเทียบระหว่างข่ายงานเบย์ (Bayesian Network) กับโมเดลแบ ANN และ SVM สำหรับระยะเวลาในการผลิต (Production Time) ในอุตสาหกรรมเหล็กกล้า พบว่า ทุกวิธีที่ใช้ในการทำนายสามารถทำนายผลลัพธ์ได้อย่างแม่นยำหากเก็บข้อมูล observed variables มาอย่างสมบูรณ์ แต่อย่างไรก็ตาม หากเราทราบ input variable ของข้อมูลมาเพียงบางส่วนเท่านั้น กลับพบว่า ข่ายงานเบย์จะให้ประสิทธิภาพที่ดีที่สุด ในกรณีนี้ซึ่งข้อมูลมาจากการจำลอง และถูกทำให้เป็นข้อมูลแบบไบนารี (Binary)

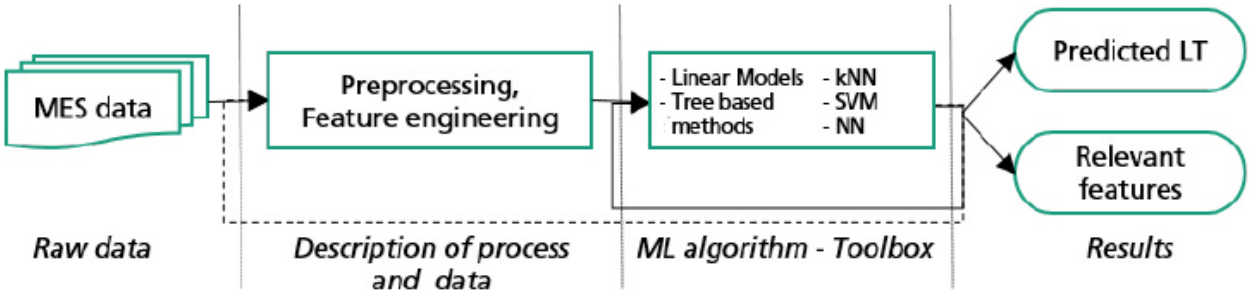
ดีคอส จูเอส และคณะ (De Cos Juez et al.) ได้ทำการวิเคราะห์ผลลัพธ์ของโมเดล SVM ที่ใช้ฟีเจอร์ทั้งหมด 8 ฟีเจอร์ (ปรับลดมาจาก 12 ฟีเจอร์) เพื่อทำการทำนายว่าชุดผลิต (batch) นี้จะทำการผลิตเสร็จสิ้นภายในระยะที่ได้ทำนาย (Forecast) ไว้หรือไม่

ลี และคณะ (Li et al.) ได้ทำการเลือกใช้โมเดลการถดถอยแบบเป็นขั้นตอน (Stepwise Regression) เพื่อที่จะประมาณค่าความสัมพันธ์ระหว่างลักษณะจำเพาะของการกระจายตัว (Distribution) ของระยะเวลาการไหล (Flow time) และตัวแปรทำนาย (Predictor variable)

เรย์เมเกอร์ส (Raaymakers et al.) พบว่าโมเดล ANN แสดงประสิทธิภาพได้ดีกว่าเล็กน้อยเมื่อเทียบกับโมเดลการถดถอย (Regression Model) ในการประมาณค่าระยะเวลาทำงาน (Make span of Job sets) ในอุตสาหกรรมการผลิตแบบชุด (bath process industry)

3. ระเบียบวิธีวิจัย (Research Methodology)

ระเบียบวิธีในการดำเนินการ ในแต่ละกรณีศึกษานั้นถูกกำหนดทิศทางด้วยรูปแบบ CRISP-DM ด้วยการเจาะจงไปที่ 5 เฟสแรก โดยที่ในส่วนของ ความเข้าใจในธุรกิจ, ความเข้าใจในข้อมูล การจัดเตรียมข้อมูล ถูกนำมารวบรวมกันอยู่ในส่วนที่เรียกว่า การอธิบายกระบวนการและข้อมูล (Description of process and data) และ 2 เฟสที่เหลือ นั่นคือ การสร้างโมเดล และการประเมินผล ถูกนำมารวมกันให้กลายเป็นส่วนใหม่ที่ชื่อว่า ML algorithm – Toolbox ดังแสดงในรูปที่ 1 โดยที่ฟีเจอร์ที่เกี่ยวข้องนั้นได้ถูกเลือก และความแม่นยำของการทำนายจากการเลือกดังกล่าวนี้ก็ได้ถูกประเมินประสิทธิภาพด้วย

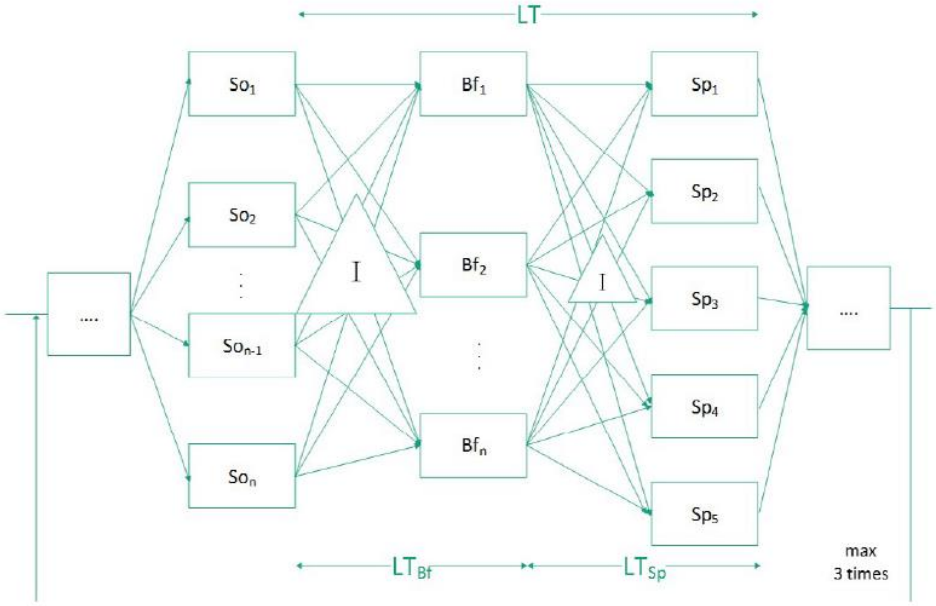


รูปที่ 1 : ลำดับขั้นในกระบวนการวิเคราะห์ข้อมูลที่นำมาใช้กับการวางแผนและควบคุมการผลิต (PPC)

3.1 การอธิบายกระบวนการผลิตและข้อมูล (Description of Manufacturing process and data)

ตามที่ได้กล่าวเอาไว้ก่อนหน้านี้ว่า กรณีศึกษานี้ได้ดำเนินการในอุตสาหกรรมเซมิคอนดักเตอร์ ซึ่งลักษณะพิเศษของอุตสาหกรรมนี้คือ ผลิตภัณฑ์ส่วนใหญ่ถูกสร้างขึ้นโดยซ้อนกันหลายชั้น (Multiple layer) นอกจากนี้ ในโรงงานผลิตแต่ละโรง (plant) ก็ไม่ได้มีการเชื่อมต่อเครือข่ายเครื่องจักรด้วยกันหมด แต่ถูกบริหารจัดการแยกโรงใครโรงมัน ส่งผลให้ระบบการผลิตแบบไม่ต่อเนื่อง (Job Shop Production System) รูปแบบนี้พบได้ว่าในหลายครั้ง ผลิตภัณฑ์เดียวกันก็ถูกผลิตจากเครื่องจักรเดียวกันเพื่อที่จะสร้างแผ่นวงจรรวม (Integrated Circuit ; IC) ในแต่ละชั้น (layer) ซึ่งแม้ว่าการจำแนกประเภทของชนิดการผลิตดังกล่าวถือว่าส่วนมากจัดเป็นแบบ Mass Production แต่ตามปกติแล้ว อีกหลากหลายผลิตภัณฑ์ก็มี lot size ที่เล็กกว่า ซึ่งเมื่อดูตามคุณสมบัติดังกล่าวของผลิตภัณฑ์ใหม่ (New Product) นั้น ๆ ก็อาจมีความจำเป็นต้องทำการผลิตโดยใช้เครื่องจักรตัวเดียวกัน จากความซับซ้อนที่ได้อธิบายไปนั้น อุตสาหกรรมเซมิคอนดักเตอร์ดั้งเดิมถึงได้ทำการลงทุนระบบ IT ในพื้นที่การผลิต ซึ่งส่งผลให้มีข้อมูลจำนวนมหาศาลเกี่ยวกับผลิตภัณฑ์ กระบวนการและเครื่องมือที่ใช้ในการผลิต ซึ่งสำหรับบริษัทภายใต้การศึกษานี้ ถูกบ่งชี้ลักษณะด้วยคุณลักษณะคือ ภายในกรณีศึกษาของแต่ละบริษัทนั้น เราได้มุ่งเน้นไปที่ลำดับ (Sequence) ของระดับขั้นกระบวนการผลิต (Process step) ทั้ง 3 ระดับ อันได้แก่ (1) **Sorter** : เป็นตัวเริ่มต้นของลำดับกระบวนการ (Process sequence) โดยจะทำหน้าที่นำแผ่นเวเฟอร์ (wafer) เหล่านั้นมาจัดลำดับให้ถูกต้องก่อนนำไปเข้าสู่ลำดับกระบวนการผลิตถัดไปคือ **(2) Breakfuse** : กระบวนการย่อยนี้อยู่ในฐานะพื้นที่กั้นกลาง (buffer) โดยมีหน้าที่ในการเตรียมแผ่นเวเฟอร์เพื่อไปเข้าสู่กระบวนการต่อไปนั่นคือ **(3) Sputter :** ในกระบวนการนี้ ชั้นแผ่นโลหะบางๆ ของแผ่นเวเฟอร์จะถูกพ่นเคลือบด้วยสารเคมีบางชนิด

ซึ่งกระบวนการเหล่านี้มีความสัมพันธ์กับค่า lay time สูงสุด (maximum lay time) ระหว่างกระบวนการ Bakefuse และ Sputter ซึ่งจำกัดขนาดของพื้นที่กั้นกลาง (buffer size) และนำไปสู่การถูกปิดทาง จากลอตที่เคลื่อนที่ไปก่อนหน้า ดังนั้นจะเห็นได้ว่า การวัด (measurement) และภายหลังนั้นได้รวมถึงการทำนายค่า lead time นั้น ไม่เพียงแต่เกี่ยวข้องกับแต่ละลอตการผลิต แต่ยังสัมพันธ์กับแต่ละระดับชั้นการผลิต (layer) อีกด้วย ซึ่งนั่นก็หมายความว่าแต่ละลอตการผลิตนั้นสามารถใช้เวลาแต่ละกระบวนการได้อย่างหลากหลาย ขึ้นกับทั้งโครงสร้างของผลิตภัณฑ์ เวลาสังเกตการณ์อาจได้มากสุดถึง 3 เท่าของเวลาปกติ เนื่องจากมันมีถึง 3 ชั้นการผลิตในลำดับการผลิตที่ทำการสังเกตนี้ ภายในระยะเวลาหลายวันที่ทำการสังเกต



รูปที่ 2 : กระบวนการผลิตซึ่งถูกนำมาวิเคราะห์

ข้อมูลยืนยันกระบวนการจากระบบ MES ซึ่งคือข้อมูล Historical data ที่เกี่ยวข้องกับสถานะอุปกรณ์ และเครื่องจักร รวมถึงข้อมูลที่เกี่ยวข้องกับลูกค้าที่ได้มาจากทางบริษัทเองในช่วงระยะเวลาย้อนหลัง 2 ปี โดยสามารถแสดงลักษณะของข้อมูลรับเข้าซึ่งถูกสรุปมาแล้วได้ดังตารางที่ 1 โดยใน 2 คอลัมน์แรกนั้นจะเป็นค่าของชื่อ และ ชนิดของข้อมูล ตามลำดับ สำหรับในคอลัมน์ที่ 3 นั้นจะพูดถึงพิสัย (range) ของตัวแปรแต่ละตัว ซึ่งจากข้อมูลที่ให้มา เราจะเห็นได้ว่าบริษัทได้ดูแลลูกค้าอยู่ 33 รายการ โดยแยกเป็น 106 ผลิตภัณฑ์และได้รับการสนับสนุนการผลิตด้วยเครื่องมือทั้งสิ้น 48 ชิ้น และหน่วยผลิตย่อย (routing) ทั้งสิ้น 38 หน่วย ระหว่างการวิเคราะห์กระบวนการทั้ง 3 กระบวนการย่อยนี้ยังประกอบไปด้วยการดำเนินการ (Operations) ทั้ง 14 รูปแบบซึ่งสามารถแยกความแตกต่างกันได้โดยอาศัยค่าระยะเวลาลงบันทึก (Time Stamp) ที่แตกต่างกัน ซึ่งประกอบไปด้วย ค่าจากการขยับเข้า (move-in) และการขยับออก (move-out) ที่แตกต่างหลากหลายรูปแบบ โดยพบความเที่ยงตรง (Precision) ของการดำเนินการในระดับวินาที (Second) โดยกราฟเส้นของค่า time stamp ของข้อมูลนี้ แสดงได้ดังรูปที่ 3 ในกรณีที่การวิเคราะห์การการจัดสรรเครื่องจักรเพื่อใช้งาน ซึ่งเรามพยายามมุ่งเน้นไปที่การสังเกตเครื่องจักรทั้ง 5 ตัว ณ กระบวนการย่อย Sputter แสดงได้ดังรูปที่ 2

ตารางที่ 1 คำอธิบายข้อมูลดิบ

Name	Type	Granularity
Product number	Alphanumeric string	106
Customer	Alphanumeric string	33
Production lot	Alphanumeric string	23819
Operations	Alphanumeric string	14
Routings	Alphanumeric string	38
Time stamp	Date and time	Seconds
Production quantity	Integer	0-25
Equipment	Alphanumeric string	43
Priority	Integer	3
Status of operations	Alphanumeric string	22

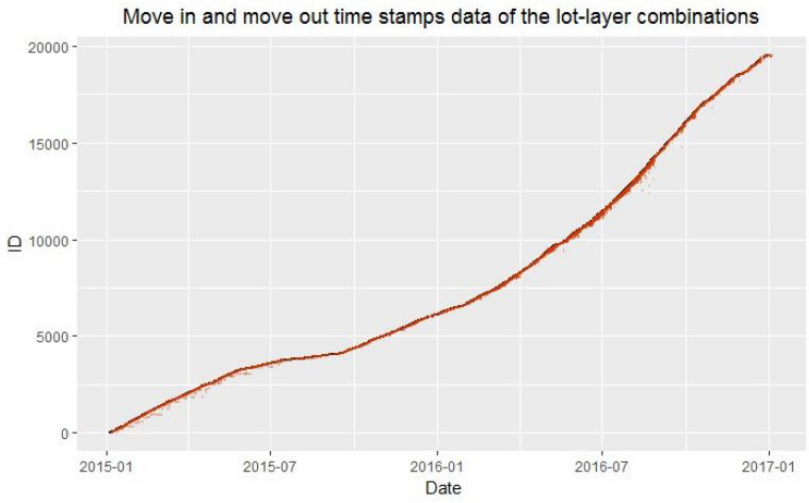
สำหรับข้อมูลการยืนยันกระบวนการที่เกี่ยวข้องกับสถานะเครื่องจักร และลูกค้านั้นได้ถูกใช้สำหรับนำมาสร้าง ML Database ซึ่งมีค่าสังเกต (Observations) ซึ่งเกี่ยวข้องกับข้อมูลรวมลอต-กระบวนการย่อย (Lot-layer combination) จำนวนทั้งสิ้น 18,532 รายการ ส่วนการประเมินประสิทธิภาพนั้น ชุดข้อมูล ML ถูกแยกออกแบบสุ่มออกเป็น training dataset และ testing dataset โดยใช้อัตราส่วนการสุ่มตัวอย่าง 70/30 ซึ่งจากชุดข้อมูลดิบที่ได้มาตอนต้นนั้น เราจะสามารถลดฟีเจอร์ลงได้เหลือเพียง 41 ฟีเจอร์ (โดยแบ่งเป็นฟีเจอร์แบบ Numerical 35 ฟีเจอร์ และแบบ Categorical อีก 6 ฟีเจอร์) ซึ่งฟีเจอร์เหล่านี้จะถูกนำมาใช้วิเคราะห์ว่าจะส่งผลอย่างไรกับกับค่า lead time และฟีเจอร์เหล่านี้ยังสามารถแบ่งประเภทโดยอิงจากลักษณะเฉพาะได้เป็น 2 รูปแบบด้วยกัน โดยแบ่งเป็น

1. ฟีเจอร์แบบสแตติก (static feature) ฟีเจอร์นี้จะเป็นตัวบ่งชี้ลักษณะเฉพาะของแต่ละลอต เช่น product , customer planned time

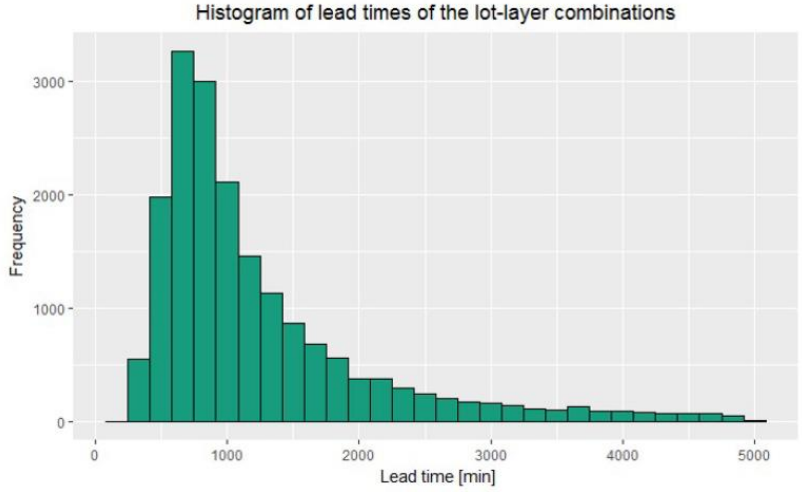
2. ฟีเจอร์แบบไดนามิค (dynamic feature) ฟีเจอร์นี้จะสะท้อนค่าสถาะของระบบการผลิตออกมาให้เราเห็น โดยเฉพาะใน 3 กระบวนการย่อนที่เราทำกาศึกษา ที่เคยกล่าวไว้ก่อนหน้านี้ ซึ่งจะทำการศึกษา ณ จุดเวลาเริ่มต้นการทำงานในแต่ละกระบวนการย่อย ของแต่ละลอต

ค่า lead time นั้นมีการคำนวณมาจากกระบวนการย่อย Bakefuse และ Sputter แยกกัน โดยสำหรับกระบวนการเหล่านี้ ค่า lead time จะถูกนิยามในฐานะช่วงเวลาจาก จุดสิ้นสุด (End Confirmation) ของกระบวนการก่อนหน้าจนถึงจุดเวลาเดียวกัน ณ กระบวนการที่ทำการสังเกต อาทิ ค่า lead time ของ Bakefuse นั้นจะถูกคำนวณมาจากค่า วัน-เวลา สิ้นสุด (end confirmation date & time) ลบออกด้วย ค่า วัน-เวลาสิ้นสุดของกระบวนการก่อนหน้า ซึ่งคือกระบวนการ Sorter ส่วนสำหรับ Spotter นั้น กระบวนการก่อนหน้าก็คือ Bakefuse

สำหรับค่า lead time รวมสำหรับ 3 กระบวนการย่อยนั้นจะคำนวณมาจากจุดสิ้นสุด (End confirmation) ของกระบวนการย่อย Sputter และจุดสิ้นสุดของกระบวนการย่อย Sorter ซึ่งโดยการอ้างอิงความรู้จากผู้เชี่ยวชาญจากในบริษัทที่ทำการศึกษานี้แล้ว งาน WIP ก่อนเข้าสู่กระบวนการ Sorter จะไม่ถูกนำมาคิดด้วย อันเนื่องมาจากเหตุผลในเชิงเทคนิคบางประการ โดยค่า lead time รวมที่ได้จากการคำนวณแสดงได้ตามรูปที่ 4



รูปที่ 3 ค่าบันทึกเวลา (time stamp) ของข้อมูลรวมลอต-กระบวนการย่อย (Lot-layer combination)



รูปที่ 4 แผนภาพฮิสโทแกรมแสดงค่า lead time ของกระบวนการที่ทำการวิเคราะห์

3.2 การสำรวจอัลกอริทึม ML (Exploring ML Algorithms)

โดยพื้นฐานแล้ว วิชาสถิตินั้นจะเกี่ยวข้องกับวิธีการต่าง ๆ ในการทำความเข้าใจเข้ามูล การค้นหาความสัมพันธ์ระหว่างพารามิเตอร์แต่ละค่า เพื่อที่นำสิ่งที่ทราบมาหาค่าประมาณของตัวแปรที่ต้องการ โดยในกรณีของการถดถอย (Regression) นั้น ค่าของตัวแปรต่อเนื่อง (Continuous variable) เป็นค่าที่เราจะทำการทำนาย ซึ่งวิธีการพื้นฐานส่วนใหญ่ที่เราจะใช้จะเป็นการใช้โมเดลแบบถดถอยเชิงเส้นแบบแปลความได้ง่าย (easily interpretable linear regression model ; LM) ซึ่งอยู่ภายใต้สมมติฐานที่จะมีตัวแปรมีความสัมพันธ์กันแบบ linear relationship โดยโมเดลกลุ่มนี้จำเป็นที่จะต้องหาค่าประมาณของพารามิเตอร์ที่จะส่งผลให้ผลรวมของค่าความคาดเคลื่อนกำลังสองเฉลี่ย (Sum of Mean Squared Error) มีค่าต่ำที่สุด ในขณะที่การถอถอยเชิงเส้นแบบธรรมดา (Ordinary Linear Regression) นั้นจะมุ่งเน้นไปที่ค่าไบแอส (Bias) ของโมเดล แต่การถดถอยแบบ Ridge และ Lasso มุ่งไปที่ค่าความแปรปรวน (Variance) ของโมเดล นอกจากนั้นแล้ว ยังมีโมเดลการถดถอยอีกหลายแบบที่โดยทั่วไปแล้วจะมีลักษณะแบบไม่เป็นเชิงเส้น (Nonlinear) เช่น ANN , Multivariate Adaptive Regression Splines (MARS) , SVM , k-nearest neighbor (KNN) และ tree based model โดยที่

1. ANN : ถือเป็นเทคนิคที่ได้รับแรงบัลดาลใจมาจากการทำงานของสมองมนุษย์ โดย Output variable นั้นถูกสร้างขึ้นมาโดยตัวแปรที่ไม่ได้ทำการสังเกต (unobserved variables) ซึ่งเราจะเรียกกลไกตรงบริเวณนี้ว่าชั้นซ่อน (Hidden layer)

2. MARS : ถือเป็นวิธีการประมาณการถดถอยแบบไร้พารามิเตอร์ (Non-parametric Regression) และสามารถนำไปใช้กับการโมเดลที่มีความสัมพันธ์แบบไม่เป็นเส้นตรงที่มีความซับซ้อนได้ รวมถึงนำไปใช้หาความสัมพันธ์ระหว่างตัวแปรได้เช่นกัน

3. SVMs : ซึ่งมีแนวทางหลักในการใช้งานนั้นจะไปใช้กับทั้งการจำแนก (Classification) และการถดถอย (Regression) นั้น จะทำการสร้างเส้นตรงในการแบ่งกลุ่มข้อมูล (hyperplane) เพื่อที่จะทำให้เกิดค่าความคาดเคลื่อน (Error) ต่ำที่สุด

4. kNN : วิธีนี้จะทำนายผลลัพธ์โดยใช้ตัวอย่างที่ใกล้ที่สุดจำนวน k ตัวกับจุดที่ต้องการทำนาย แล้วดูผลลัพธ์ว่ามีความใกล้เคียงกับข้อมูลตรงส่วนใดมากที่สุด

5. Regression Tree ; RT : เป็นการนำเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มาใช้เพื่อแก้ไขปัญหาแบบ Regression ซึ่งแม้ว่าวิธีการนี้ตีความและใช้งานได้ง่าย แต่ก็ยังมีข้อจำกัดตรงที่ประสิทธิภาพในการทำนายก็อาจจะน้อยกว่าที่ควร และรวมทั้งยังมีค่าไม่แน่นอน

6. Ensemble Technique: เป็นเทคนิคที่จะใช้โมเดลการจำแนก (Classification) หลายโมเดลมาช่วยในการหาคำตอบเพื่อลดค่าความแปรปรวน (Variance) ของการทำนายและเพื่อเพิ่มค่า Accuracy และ Stability ให้กับโมเดล แต่ประสิทธิภาพที่มีการเพิ่มขึ้นนั้น ก็ต้องมีสิ่งที่แลกไปเหมือนกัน อาทิ ค่าใช้จ่ายที่ใช้ในการรันโมเดล (Computational Cost) , ความต้องการของพื้นที่บนหน่วยความจำ (Memory Requirement) หรือการสูญเสียข้อมูลในขณะทำการแปรผล (Loss in Interpretation) (ซึ่งวิธีการ Ensemble นั้น สามารถนำ เทคนิค RT และ Bagging & Boosting มาใช้ในการหาผลลัพธ์ร่วมกัน เพื่อเพิ่มประสิทธิภาพในการทำนายได้เช่นกัน)

7. Bagging & Boosting : เป็นวิธีการ Ensemble ทั่วไปสองวิธีที่ช่วยปรับปรุงความแม่นยำในการทำนาย โดยที่ Bagging (Bootstrap Aggregation) เป็นวิธีการทั่วไปที่ใช้ bootstrapping มาช่วยในการทำงาน โดยการทำนายของ Bagged Regression Tree Model นั้นจะมีผลลัพธ์ออกมาเป็นค่าเฉลี่ยของการทำนายระว่างแต่ละ Regression Tree ใน Bagged Ensemble

9. Random Forest (RF) : รูปแบบนี้ถือว่าเป็นกรณีพิเศษของ Bagged Regression Tree ซึ่งในวิธีนี้จะทำการสุ่มเลือกตัวแปรทำนาย (Predictor) มาใช้ในกระบวนการสร้างต้นไม้ (Tree) เพื่อนที่จะลดการเกิดความสัมพันธ์กันเอง (Correlation) จากตัวแปรทำนาย ซึ่งจุดที่ทำให้ RF แตกต่างนั่นคือ RF นั้น ต้นไม้แต่ละต้นจะเป็นอิสระต่อกัน ต้นไม้ทุกต้นจะมีสิทธิ์เท่าเทียมกันในการร่วมกันสร้างโมเดลสุดท้าย และต้นไม้แต่ละต้นสามารถที่จะถูกสร้างไปถึงค่าความลึกสูงสุด (maximum depth) ในขณะที่ Boosting นั้น ต้นไม้จะไม่เป็นอิสระต่อต้นไม้ตัวก่อนหน้า มีรูปแบบการทำงานของการสร้างต้นไม้หลากหลายรูปแบบ และไปถึงแค่ความลึกต่ำสุด (Minimum Depth)

3.3 การประยุกต์ใช้และการประเมินประสิทธิภาพของเทคนิค ML ที่เลือกใช้

วิธีการเรียนรู้ทางสถิติทั้ง 11 ชนิดนั้นถูกประเมินประสิทธิภาพการในการวิเคราะห์ โดยใช้ภาษา R ด้วย R Studio ในกรณีของ linear model (LM , Ridge และ Lasso) และ ANN Regression นั้น เฉพาะ Numerical Feature เท่านั้นที่จะถูกนำมาใช้งาน โดยระหว่างกระบวนการสร้างโมเดลและการเลือกฟีเจอร์นั้น 10-fold cross validation จะถูกนำมาใช้เพื่อประเมินค่าความแม่นยำในกาทำนายของโมเดลที่มีต่อชุดข้อมูลฝึกฝนแบบอิสระ โดยความแม่นยำของโมเดลนั้นจะถูกวัดผลด้วยรูปแบบการวัดค่าความคาดเคลื่อน (Error) ทั้ง 5 รูปแบบ อันได้แก่ Mean Absolute Error (MAE) , Mean Absolute Percentage Error (MAPE) , Mean Squared Error (MSE) , Root Mean Squared Error (RMSE) และ Normalized Root Mean Squared Error (NRMSE) โดยสำหรับ NRMSE นั้นจะให้ค่าเฉลี่ยของการทำนายความคาดเคลื่อนในรูปของเปอร์เซ็นต์ของค่า lead time จริง ซึ่งในการคำนวณค่า NRMSE นั้น เป็นไปตามสมการที่ 1 โดยที่ P_i และ R_i นั้นคือค่าที่เกิดจากการทำนาย และค่าจริงของ lead time ตามลำดับ โดยผลลัพธ์ของโมเดลเหล่านี้สามารถสรุปได้ตามตารางที่ 2

$$NRMSE = 100\frac{\sqrt[2]{\frac{1}{n}\sum_{i=1}^{N}(P_i - R_i)}}{R_{max} - R_{min}}$$

.....(1)

ซึ่งเหล่านี้สามารถสรุปได้ว่า Linear Model ทั้ง 3 ตัวให้ค่าความแม่นยำ (Accuracy) เกือบจะเหมือนกัน (ลองพิจารณาการวัดค่าคาดเคลื่อน จากตารางที่ 2 ประกอบ) ซึ่งแต่ละโมเดลมีการตั้งค่าพารามิเตอร์ตามรายละเอียดดังนี้

1. RF model ผลลัพธ์ของโมเดล คำนวณมาจากค่า Default parameter ซึ่งประกอบด้วย จำนวนต้นไม้ (tree) = 500 , จำนวนตัวแปรสุ่มที่เลือก (random variable) = 13 ตัว จากตัวแปรทั้งหมดที่มี 41 ตัว , โดยจากค่าที่แนะนำนั้น จะใช้ปริมาณตัวแปรทำนาย (Predictor) ประมาณ 1/3 ของตัวแปรทั้งหมดในกรณีของ Regression

2. Boosted RT Model นั้นประกอบด้วย tree = 20,000

3. SVM นั้นมีการใช้งาน Radial Kernel

4. kNN จากการทดลองพบว่าค่า k ที่เหมาะสมคือ k = 9 ภายหลังการทดสอบค่า k ตั้งแต่ k=5 ถึง k = 23

5. ANN model ประกอบด้วย Hidden Layer = 1 และ Neuron = 26 ซึ่งสำหรับค่าที่แนะนำคือ จำนวน Input variable ลดลงเหลือเพียง 2/3 ของทั้งหมด (ซึ่งนั่นก็จะทำให้ค่า NRMSE นั้นจะพบที่การปรับค่า neuron ระหว่าง 20 - 30)

และจากผลลัพธ์ที่สรุปไว้ในตารางที่ 2 จะเห็นว่าวิธีการแบบ Ensemble Tree Based นั้นทุกโมเดลสามารถแสดงความสามารถได้อย่างโดดเด่น แต่โมเดลที่ผลลัพธ์ดีที่สุดก็ยังคงเป็น RF โดยทั้ง Bagging และ RF นั้นต่างก็ถูกสร้างขึ้นมาด้วย*แพคเกจ RandomForest* ส่วนสำหรับ Boosted RT นั้นถูกสร้างขึ้นมาด้วยแพคเกจ gbm โดยธรรมชาติ การรวมกันของโมเดลพวกนี้นั้นเป็นไปไม่ได้เลยที่จะทำให้เราเข้าใจความสัมพันธ์ระหว่าง input และ output variable แต่อย่างไรก็ดีก็ยังมีความเป็นไปได้ในการกำหนดปริมาณผลกระทบของตัวแปรทำนายใน ensemble โดยค่าฟังก์ชั่นสำคัญ (Importance Function) ที่แพคเกจ RandomForest นั้นสามารถประเมินระดับความสำคัญของทุกตัวแปรได้ด้วย 2 วิธีการดังต่อไปนี้

i) การเรียงสับเปลี่ยนแบบสุ่มของค่าของแต่ละตัวแปรทำนาย (Random permutation of the values of each predictor) หรือ

ii) การปรับปรุงความบริสุทธิ์ของโหนดโดยอิงจากตัวชี้วัดประสิทธิภาพของแต่ละตัวแปรทำนาย (Improvement in node purity based on the performance metric of each predictor)

ความสำคัญของตัวแปรสำหรับการบูสต์ นั้นคือฟังก์ชั่นการลดในรูปความคาดเคลื่อนกำลังสองอันเนื่องมาจากตัวแปรทำนายแต่ละตัว โดยจุดตัดสำหรับตัวแปรที่มีความสำคัญสูงสุด ซึ่งพบโดยแพคเกจ *RandomForest* และ *gbm* นั้นถูกสรุปและอธิบายอยู่ในสองคอลัมน์แรก ณ ตารางที่ 3 ด้วยความช่วยเหลือของ 10 ตัวแปรที่มีความสำคัญสูงสุดนี้เอง โมเดล RF และ boost RT จึงถูกสร้างขึ้นมา ในฐานะข้อแนะนำสำหรับการทำนาย lead time ซึ่งระหว่างการปรับค่าพารามิเตอร์ในโมเดลแบบละเอียด (fine tuning) ก็ได้ค้นพบว่า ค่าจำนวนของต้นไม้ (Number of Trees) ที่เหมาะสมในการสร้างโมเดล RF นั้นจะอยู่ในช่วง 100-125 โดยที่การทดลองจะเริ่มต้นที่ Tree = 500 (running time = 2 นาที) จะได้ค่า NRMSE มีค่าเป็น 13.1 ซึ่งหากลองลดค่า Tree ลงอีกเป็น Tree = 125 (running time < 1 นาที) ค่า NRMSE ก็ยังมิได้มีการเพิ่มขึ้น จนเมื่อลองลด Tree ลงอีกเป็น Tree = 100 (running time 20 - 30 s) จะพบว่า ค่า NRMSE นั้นเพิ่มขึ้นมาจากเดิมอยู่ 0.1 (กลายเป็น 13.2)

สำหรับค่าระยะเวลาการรันสำหรับโมเดล Boosted RT นั้นจะมีนัยสำคัญมากกว่า (เริ่มต้นที่ running time = 6 นาที , Tree = 25,000 และ running time = 5 นาที , Tree = 20,000) และค่าความแม่นยำ (Accuracy) ของโมเดลนี้มิได้โดดเด่นเหนือไปกว่าผลลัพธ์ที่ได้จากโมเดล RF (ได้ค่า NRMSE ของการรันทั้ง 2 มีค่าเป็น 13.5 และ 13.6 ตามลำดับ)

ภายหลังกระบวนการสร้างโมเดล RF และ Boosted RT ขั้นสุดท้ายแล้วนั้น ค่าระดับความสำคัญของตัวแปรทั้ง 10 นั้นได้มีการถูกนำมาศึกษาด้วยความช่วยเหลือของการวิเคราะห์ความอ่อนไหว (Sensitivity Analysis) สำหรับผลการวิเคราะห์ความอ่อนไหวนั้นได้ถูกสรุปอยู่ ณ 2 คอลัมน์สุดท้ายในตารางที่ 3 ซึ่งจะพบว่าค่า MSE (ในรูป %) นั้นมีค่าเพิ่มขึ้น โดยในการวิเคราะห์นี้ ยังมิได้ทำการวิเคราะห์อย่างเจาะจงเป็นพิเศษกับตัวแปรใดๆ (ในกรณีของโมเดล RF ผลลัพธ์ของการรันโมเดล 5 ครั้ง ถูกเฉลี่ยออกมา) ซึ่งหากอ้างอิงตามผลลัพธ์ที่ได้นี้ จำนวนของ lot-layer ในกระบวนการย่อยที่ทำการวิเคราะห์ (WIP) จะปรากฏออกมาเพื่อที่จะเป็นตัวแปรที่มีระดับความสำคัญสูงสุด ซึ่งการมองข้ามฟีเจอร์นี้เมื่อถึงโมเดลสุดท้าย ก็จะมีผลกระทบมากที่สุดอย่างมีนัยสำคัญต่อค่า MSE โดยการลบตัวแปร WIP ซึ่งอยู่ในฐานะฟีเจอร์ของ Final RF Model และ Boosted RT Model นั้น ส่งผลให้ค่า MSE ของทั้งสองโมเดลถูกเพิ่มขึ้นเป็น 5.3% และ 2.6% ตามลำดับ และหากหากพิจารณาค่าผลลัพธ์การวิเคราะห์ความอ่อนไหว สำหรับตัวแปร 2 ตัว (ArriveHour และ MovArrival) มีค่าเป็นลบ ดังนั้นในกรณีนี้ผลลัพธ์ได้จะออกมาดีกว่า หากเราทำการสร้างโมเดล โดยตัดฟีเจอร์ 2 ตัวนี้ออกไป

หากยึดตามข้อมูลที่ได้อ้างไว้ในหนังสือ Applied Predictive Modeling ซึ่งแต่งโดย Max Kuhn และKjell Johnson แล้วนั้น เราก็จะพบว่ามีค่าไบแอส (Bias) ในการวัดค่าระดับความสำคัญของตัวแปรในโมเดล RF มีตัวอย่างอยู่ 2 ตัวอย่างซึ่งถือได้ว่ามีผลอย่างร้ายแรงต่อค่าระดับความสำคัญ นั่นคือ (1) ค่าสหสัมพันธ์ระหว่างตัวแปรทำนาย และ (2) จำนวนของตัวแปรสุ่มระหว่างการสร้างโมเดล ซึ่งการวิเคราะห์ในขั้นต่อไปนั้นมีความจำเป็น เพื่อทำให้เกิดความเข้าใจความสัมพันธ์ซึ่งกันและกันระหว่างตัวแปรทั้ง 2 ตัว ตามที่ได้กล่าวไว้ข้างต้น และยังรวมถึงตัวแปรอื่นๆ ด้วย

จากการศึกษานี้ โมเดลสุดท้ายที่เราแนะนำให้ใช้ในการทำนาย lead time นั้น คือโมเดล RF ที่ใช้ตัวแปร 8 ตัวในการทำนาย และผลการวิเคราะห์ค่าความอ่อนไหวของตัวแปรที่เลือกเหล่านั้นเป็นบวก ดังแสดงในตารางที่ 3 ด้วยพารามิเตอร์คือ tree = 125 และใช้ Random Variable = 2 ในการสร้างโมเดล ซึ่งจะใช้ Running time 20 วินาที (ลดมาจากโมเดลตั้งต้นซึ่งมีถึง 41 ฟีเจอร์) โดยผลลัพธ์สุดท้ายที่ได้ คือ ได้ NRMSE เป็น 12.5

ตารางที่ 2 ค่าคาดเคลื่อน (Error) รูปแบบต่าง ๆ ของโมเดลที่ถูกนำมาทดสอบ

	LM	Ridge	Lasso	RT	bagged RT	RF	boosted RT	SVM	MARS	kNN	ANN
MAE	487	510	508	563	394	390	397	423	488	504	535
MAPE	42.7	45.0	44.7	53.5	33.9	33.8	33.9	30.9	43.15	44.0	53.4
MSE	529408	573520	572939	639617	369993	360780	369414	500693	513638	554897	658852
RMSE	727	757	756	799	608	600	607	707	716	745	771
NRMSE	15.2	15.8	15.8	16.7	12.7	12.5	12.7	14.77	14.9	15.5	17.3

ตารางที่ 3 คำอธิบายและผลการทดสอบค่าความอ่อนไหว (Sensitivity Analysis) ของตัวแปรที่สำคัญที่สุด 10 อันดับแรกของโมเดล

Feature	Description	RF	boosted RT
MovDeparture	Moving average of the inter departure times of the last 20 lot-layers	2,7	3.3
ArrivalHour	Hour of the arrival time	-7,6	-13.9
WD	Weekday of the arrival time	3.6	2.7
SumMedOTs	Median of the product‘s lead time (in the analyzed process steps)	5.2	1.5
WIP	Work in progress: number of lot-layers in the analyzed process steps	5.3	2.6
WIPtimeBfMed	Work in progress: expected work content in minutes by process step bakefuse	2.9	1.7
SpEffPrevDay	Capacity utilization of machines in process step sputter on the previous day	0.1	1.9
MovArrival	Moving average of the inter arrival times of the last 10 lot-layers	-1.3	-1.3
medOTProdRout	Mean of median operations times of a product on a given route	3.6	2.1
SBPrevDay	Time in standby status of the machines of sputter on the previous day	1.2	0.8

4. หัวข้องานวิจัยในอนาคต (Future Research Agenda)

การวิจัยต่อจากจุดนี้จะถูกดำเนินการต่อโดยผู้แต่ง โดยในขั้นแรกนั้น ขอบเขตของการวิเคราะห์จะถูกขยายผลและวิธีการที่ถูกพัฒนาขึ้นแล้วจะถูกนำไปใช้กับกระบวนการย่อยอื่น ๆ รวมถึงตลอดทั้งระบบการผลิตนี้ เพื่อทำการวิเคราะห์ ซึ่งถ้าวิธีการที่นำมาใช้ยังคงมีความเหมาะสมสำหรับส่วนอื่น ๆ แล้ว หัวข้องานวิจัยหลังจากนี้ก็จะเป็นการประยุกต์ใช้วิธีการนี้กับอุตสาหกรรมอื่น ๆ เพื่อที่จะตรวจสอบความเหมาะสมของตัวแปรและอัลกอริทึมที่ใช้

จากการทบทวนวรรณกรรมและงานวิจัยที่ได้ดำเนินการนี้ ความจำเป็นของ Feature Codebook ก็ได้ปรากฏขึ้น ดังนั้น ฟีเจอร์ใหม่จำเป็นที่จะต้องถูกนิยามความหมายและทดสอบ รวมถึงฟีเจอร์ที่มีอยู่แล้วจำเป็นที่จะต้องปรับค่า (tune) อีกทั้งประโยชน์ใช้สอย และความเหมาะสมของฟีเจอร์เหล่านั้นที่มีต่อชนิดของกระบวนการผลิตที่แตกต่างกันไป (อาทิ กระบวนการแบบไม่ต่อเนื่อง (batch process) , กระบวนการผลิตแบบต่อเนื่อง (Continuous process) และรูปแบบอื่น ๆ) รวมถึงอุตสาหกรรมจำเป็นที่ต้องมีการศึกษาและจัดทำเอกสารประกอบมากยิ่งขึ้น

สุดท้าย แต่ไม่ท้ายสุด การวิเคราะห์ความสัมพันธ์ซึ่งกันและกันระหว่างตัวแปรแต่ละตัว (Interrelation between variables) และค่าความอ่อนไหวเชิงลบ (Negative Sensitivity) จะถูกสืบสวนต่อไป