

ห้องสมุดคณะเทคโนโลยีสารสนเทศ พระจอมเกล้าลาดกระบัง

ระบบช่วยวิเคราะห์ความเสี่ยงของลูกค้าโทรศัพท์เคลื่อนที่ด้วยดัชนีชี้วัด

CREDIT RISK ANALYSIS BY DECISION TREE
FOR MOBILE PHONE CUSTOMER



H005969

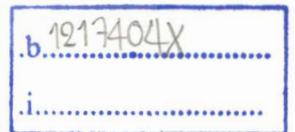


ทพ.
๘๑๙๔
๒๕๕๑

เลขหมู่.....

เลขทะเบียน.....05969

วัน,เดือน,ปี. ๓.๑.๒๕๕๓



รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับภาคเรียนที่ 2 ปีการศึกษา 2551 อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

**CREDIT RISK ANALYSIS BY DECISION TREE
FOR MOBILE PHONE CUSTOMER**



**A SYSTEM DEVELOPMENT PROJECT
OF THE REQUIREMENT FOR THE DEGREE OF
MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY
FACULTY OF INFORMATION TECNOLOGY**

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
2/2008
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



COPYRIGHT 2009

FACULTY ON INFORMATION TECHNOLOGY

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับกรณีใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้ทำไปใช้ประโยชน์ด้านการค้า
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อ	ระบบช่วยวิเคราะห์ความเสี่ยงของลูกค้าโทรศัพท์เคลื่อนที่ด้วยดิจิทัล
นักศึกษา	นายวิสุทธิ ลีลาชุดิพงศ์
รหัสนักศึกษา	47066625
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2551
อาจารย์ที่ปรึกษา	รศ.ดร.วรพจน์ กรีสระเดช

บทคัดย่อ

การบริหารลูกค้าสัมพันธ์ และการบริหารความเสี่ยงเป็นหัวข้อที่ได้รับความนิยมอย่างมาก ในธุรกิจในปัจจุบัน ซึ่งระบบช่วยวิเคราะห์ความเสี่ยงของลูกค้าโทรศัพท์เคลื่อนที่ด้วยดิจิทัลจะเป็นระบบที่ช่วยสนับสนุนข้อมูลด้านการตัดสินใจให้กับผู้ใช้งานในการวิเคราะห์ขีดความสามารถทางการเงินของลูกค้ารายใหม่ได้อย่างถูกต้อง และยังช่วยลดเวลาในการวิเคราะห์ข้อมูล ถึงแม้ว่าผู้ใช้งานนั้นจะไม่มีความรู้และประสบการณ์ด้านการวิเคราะห์ความเสี่ยงก็ตาม อีกทั้งยังช่วยลดความเสี่ยงของการเกิดหนี้ และลดต้นทุนในการติดตามหนี้ของผู้ให้บริการเครือข่ายโทรศัพท์เคลื่อนที่ด้วย

Title	Credit Risk Analysis by Decision Tree for Mobile Phone Customer
Student	Mr. Wisut Leelachutipong
Student ID.	47066625
Degree	Master of Science in Information Science
Programme	Information Science
Academic Year	2008
Advisor	Assoc.Prof.Dr. Worapoj Kreesuradej

ABSTRACT

Customer Relationship Management (CRM) and Risk Management have become significant topic for every business. Credit Risk Analysis by Decision Tree Application is a tool for user in order to classify or evaluate the credit risk of current customers and to predict the potential of new customers. Furthermore, it can decrease the bad debts and cost of collection in business as well.



กิตติกรรมประกาศ

การพัฒนาระบบงานและการศึกษาหลักสูตรนี้ ทางผู้จัดทำขอขอบพระคุณ รศ.ดร.วรพจน์ กรีสระเดช ที่เป็นอาจารย์ที่ปรึกษาและให้คำแนะนำอย่างดียิ่ง อีกทั้งคณาจารย์ทุกท่านที่ได้ให้ความรู้ซึ่งมีประโยชน์ในการทำงานเป็นอย่างยิ่ง ขอขอบคุณเพื่อนๆ IS 18.2 โดยเฉพาะเพื่อนๆในกลุ่มของข้าพเจ้าที่คอยช่วยเหลือด้วยดีตลอดมา ขอขอบคุณพี่ระดมพล พี่วิฑูร พี่ชัยสิทธิ์ พี่นริศ และเพื่อนๆ ที่ช่วยเหลือ ตักเตือน ดิชม และให้กำลังใจเสมอมา ที่สำคัญยังขอขอบพระคุณบิดา มารดา พี่ และน้องของข้าพเจ้าที่เป็นพลังอันสำคัญและคอยเป็นกำลังใจให้ข้าพเจ้าเสมอมา

วิสุทธิ์ ธิลาชุดิพงศ์



สารบัญ

	หน้า
บทคัดย่อภาษาไทย.....	I
บทคัดย่อภาษาอังกฤษ.....	II
กิตติกรรมประกาศ.....	III
สารบัญ.....	IV
สารบัญตาราง.....	VI
สารบัญรูป.....	VII
บทที่	
1. บทนำ.....	1
1.1 ความสำคัญและที่มา.....	1
1.2 วัตถุประสงค์ของโครงการ.....	2
1.3 ขอบเขตการดำเนินงาน.....	2
1.4 เครื่องมือที่ใช้ในการพัฒนาระบบ.....	2
1.5 ขั้นตอนและวิธีการดำเนินงาน.....	2
1.6 ประโยชน์ที่คาดว่าจะได้รับ.....	3
2. การบริหารลูกค้าสัมพันธ์และการประเมินพฤติกรรมด้านสินเชื่อ.....	4
2.1 การบริหารลูกค้าสัมพันธ์.....	4
2.2 การทำค้ำไ่มนึ่งเพื่การบริหารลูกค้าสัมพันธ์.....	5
2.3 การประเมินพฤติกรรมด้านสินเชื่อ.....	7
3. ค้ำไ่มนึ่ง.....	9
3.1 ความหมายของค้ำไ่มนึ่ง.....	9
3.2 ขั้นตอนการทำค้ำไ่มนึ่ง.....	9
3.3 การจัดหมวดหมู่ของข้อมูล.....	15
3.4 การประเมินค่าความแม่นยำของค่าพยากรณ์ที่ได้จากแบบจำลองพยากรณ์.....	16

สารบัญ (ต่อ)

	หน้า
4. ต้นไม้ช่วยตัดสินใจ.....	18
4.1 การสร้างต้นไม้ช่วยตัดสินใจ.....	18
4.2 ข้อจำกัดของต้นไม้ช่วยตัดสินใจ.....	19
4.3 อัลกอริทึม SLIQ.....	20
4.4 หลักการสร้างต้นไม้ตัดสินใจ.....	21
4.5 การสร้างแบบจำลองต้นไม้.....	22
4.6 การขยายความสามารถของต้นไม้ตัดสินใจ.....	27
5. การพัฒนาโปรแกรม.....	28
5.1 การศึกษาทฤษฎีที่เกี่ยวข้อง.....	28
5.2 การศึกษาความเป็นไปได้ และวิเคราะห์ออกแบบระบบ.....	28
5.3 การจัดเตรียมและจัดเก็บข้อมูลในการวิเคราะห์.....	29
5.4 การออกแบบระบบ.....	32
5.5 การใช้งานระบบช่วยวิเคราะห์ความเสี่ยงของลูกค้าโทรศัพท์เคลื่อนที่.....	50
6. บทสรุปและข้อเสนอแนะ.....	61
6.1 สรุปผล.....	61
6.2 ประโยชน์ที่ได้รับ.....	61
6.3 ข้อเสนอแนะและแนวทางในการพัฒนาโครงการเพิ่มเติม.....	62
บรรณานุกรม.....	63
ประวัติผู้เขียน.....	64

สารบัญตาราง

ตารางที่	หน้า
3.1 ผลของการประเมินแบบจำลองพยากรณ์.....	16
5.1 แอททริบิวต์ที่ใช้ในการวิเคราะห์และประเมินความเสี่ยง.....	29
5.2 แอททริบิวต์ที่เก็บผลจากการวิเคราะห์และประเมินความเสี่ยง.....	31
5.3 Use Case Description ของ Select Data	34
5.4 Use Case Description ของ Edit Parameter	35
5.5 Use Case Description ของ Build Model	36
5.6 Use Case Description ของ View Model	37
5.7 Use Case Description ของ View Classification Rule	38
5.8 Use Case Description ของ Save Model	39
5.9 Use Case Description ของ Select Model	40
5.10 Use Case Description ของ Edit Customer's profile	41
5.11 Use Case Description ของ Predict Data	42
5.12 Use Case Description ของ Save Predicted Data	43
5.13 Use Case Description ของ Search Predicted data	44
5.14 Use Case Description ของ View Predicted data table	45

สารบัญรูป

รูปที่	หน้า
3.1 กระบวนการทำค้ำไม้หนึ่งโดยทั่วไป.....	14
3.2 กระบวนการ Classification	15
4.1 อัลกอริทึมพื้นฐานในการสร้างต้นไม้ช่วยตัดสินใจ.....	18
4.2 ตัวอย่างของ Decision Tree	19
4.3 วิธีทำ Cross Validation ด้วยวิธี K-Fold.....	22
4.4 การสร้างคลาสสิกด้วยวิธีเรียงลำดับก่อน.....	23
4.5 อัลกอริทึมการแบ่งโหนดของ SLIQ	24
4.6 การแบ่งโหนดโดยใช้แอททริบิวต์ Age เป็น Splitting Index	24
4.7 การแบ่งโหนดของต้นไม้ตัดสินใจโดยใช้แอททริบิวต์ Salary เป็น Splitting Index....	25
4.8 อัลกอริทึมการปรับปรุงคลาสสิกของต้นไม้ตัดสินใจ.....	26
4.9 การแบ่งโหนดของต้นไม้ตัดสินใจย่อย.....	26
4.10 การปรับปรุงคลาสสิกของต้นไม้ตัดสินใจ.....	27
5.1 Use Case Diagram แสดงความสัมพันธ์ระหว่างผู้ใช้และฟังก์ชันต่างๆ ของระบบ.....	33
5.2 Activity Diagram แสดงการทำงานของระบบ.....	46
5.3 Class Diagram แสดงการทำงานของระบบ.....	49
5.4 หน้าจอหลักของโปรแกรม.....	50
5.5 เมนูหลัก Create Model.....	51
5.6 การเลือกข้อมูลจากฐานข้อมูล.....	52
5.7 กล่องข้อความผิดพลาดหากไม่พบข้อมูล.....	52
5.8 ตัวอย่างการใส่ค่าพารามิเตอร์เพื่อสร้างแบบจำลองแสดง.....	53
5.9 แบบจำลองที่สร้างในรูปแบบต้นไม้ตัดสินใจ.....	53
5.10 แบบจำลองที่สร้างในรูปกฎการตัดสินใจ if-then.....	54
5.11 ข้อมูลของ Training Data และการประเมินแบบจำลอง.....	54
5.12 หน้าจอเพื่อบันทึกแบบจำลองที่สร้างไว้.....	55
5.13 เมนูหลัก Data Prediction.....	55
5.14 หน้าจอเพื่อเลือกแบบจำลองที่ต้องการใช้พยากรณ์.....	56
5.15 แบบจำลองที่ถูกเลือกใช้พยากรณ์.....	56
5.16 การใส่ข้อมูลของลูกค้าที่ต้องการพยากรณ์.....	57

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น เมื่ออนุญาตให้นำไปเผยแพร่โดยไม่ได้รับอนุญาต

สารบัญรูป (ต่อ)

	หน้า
5.17 ผลการพยากรณ์ของลูกค้า.....	57
5.18 กล่องข้อความเพื่อยืนยันการบันทึกผล.....	58
5.19 กล่องข้อความเมื่อบันทึกผลสำเร็จ.....	58
5.20 เมนูหลัก Credit Report.....	58
5.21 การค้นหาข้อมูล.....	59
5.22 การเลือกเงื่อนไขที่ต้องการ.....	59
5.23 ผลลัพธ์ที่ต้องการ.....	60



บทที่ 1

บทนำ

1.1 ความสำคัญและที่มา

ในปัจจุบันธุรกิจผู้ให้บริการเครือข่ายการสื่อสารโทรคมนาคมมีสถานะการแข่งขันที่สูงมาก บริษัท หรือองค์กรต่างๆ จึงให้ความสำคัญในการสร้างฐานความรู้สำหรับการตัดสินใจเพื่อตอบสนองต่อเป้าหมายของธุรกิจ หรือการสร้างแผนกลยุทธ์เพื่อกำหนดแนวทางในการปฏิบัติงาน เช่น การสร้างโอกาสทางธุรกิจ การสร้างความพึงพอใจให้กับลูกค้า การส่งเสริมการขาย เพื่อเพิ่มรายได้ เป็นต้น สิ่งเหล่านี้จำเป็นต้องนำการจัดการสารสนเทศเข้ามาช่วยเพื่อวิเคราะห์ข้อมูล ในองค์กรที่มีปริมาณข้อมูลมากให้ได้สารสนเทศที่มีประโยชน์เพื่อทำให้ประสบความสำเร็จ หรือ บรรลุวัตถุประสงค์ตามเป้าหมายต่างๆ ที่วางไว้ ดังนั้นดาต้าไมนิ่งจึงเป็นเครื่องมือที่สำคัญใน เทคโนโลยีสารสนเทศซึ่งสามารถนำมาช่วยวิเคราะห์และค้นหาสารสนเทศที่ซ่อนอยู่ในข้อมูล เหล่านั้นได้

นอกจากเป้าหมายในการสร้างแผนกลยุทธ์ หรือการสร้างฐานความรู้เพื่อตัดสินใจที่ได้กล่าวข้างต้นแล้ว ทุกวันนี้ธุรกิจต่างๆ ก็ให้ความสำคัญของการจัดการความเสี่ยง (Risk Management) อีกด้วย เช่น การให้คะแนนความสามารถในการชำระหนี้ของลูกค้า (Credit Scoring) การประเมิน หนี้เสีย การป้องกันการฉ้อโกง (Fraudulent Activity) การหาข้อผิดพลาดของการปฏิบัติงาน หรือ การให้บริการ (Fraud Detection) เป็นต้น เพราะว่าความเสี่ยงต่างๆ เหล่านี้ทำให้เกิดค่าใช้จ่ายที่สูงมากและเป็นอุปสรรคที่ทำให้ธุรกิจไม่สามารถประสบความสำเร็จตามแผนกลยุทธ์ที่วางไว้

โครงการนี้จะนำดาต้าไมนิ่งมาประยุกต์ใช้ในการจัดการความเสี่ยงของธุรกิจสื่อสาร โทรคมนาคมด้วยแนวความคิดในการวิเคราะห์และประเมินความเสี่ยงของลูกค้า (Credit Risk Analysis) โดยการจัดหมวดหมู่ และการพยากรณ์จากข้อมูลของลูกค้าทั้งในอดีต ข้อมูลปัจจุบัน และ ข้อมูลตัวอย่าง ซึ่งทำให้สามารถจัดสารสนเทศให้เป็นหมวดหมู่ และประเมินความสามารถในการ ชำระหนี้ของลูกค้าได้อย่างเหมาะสม ทำความเข้าใจพฤติกรรมของลูกค้าในแต่ละกลุ่ม หรือการ พยากรณ์กลุ่มของประชากรที่คาดว่าจะมาเป็นลูกค้าในอนาคต ซึ่งทำให้สามารถลดต้นทุนในการ ติดตามทวงหนี้และลดการสูญเสียของรายได้ที่เกิดจากหนี้เสีย หรือการฉ้อโกง เป็นต้น โดยการ วิเคราะห์ข้อมูลจะใช้หลักการที่เรียกว่า การจัดหมวดหมู่ของข้อมูล และการพยากรณ์ข้อมูล (Classification and Prediction) แล้วนำผลของข้อมูลผ่านการวิเคราะห์นั้นมานำเสนอในรูปแบบ ของแผนภูมิต้นไม้ช่วยตัดสินใจ (Decision Tree) ซึ่งเป็นรูปแบบที่สามารถเข้าใจได้ง่าย

1.2 วัตถุประสงค์ของโครงการ

1. เพื่อศึกษาแนวคิดและขั้นตอนของการทำดาต้าไมนิ่งด้วยเทคนิคจัดหมวดหมู่ของข้อมูล (Data Classification) โดยใช้อัลกอริทึม SLIQ (Supervised Learning in Quest)
2. เพื่อศึกษาแนวคิดและนำเสนอข้อมูลจากการทำดาต้าไมนิ่งโดยนำหลักการและรูปแบบของต้นไม้ช่วยตัดสินใจ
3. เพื่อนำความรู้และเทคนิคที่ได้ศึกษาไปประยุกต์ใช้กับการวิเคราะห์ความเสี่ยงของการให้เครดิตกับลูกค้า
4. เพื่อนำความรู้และเทคนิคที่ได้ศึกษาไปประยุกต์ใช้กับการวิเคราะห์ข้อมูลทางธุรกิจผู้ให้บริการเครือข่ายสื่อสารโทรคมนาคม

1.3 ขอบเขตการดำเนินงาน

การพัฒนาระบบนี้เป็นศึกษาแนวคิดและขั้นตอนของการทำดาต้าไมนิ่งด้วยเทคนิคจัดหมวดหมู่ของข้อมูล (Data Classification) และนำเสนอข้อมูลที่ได้จากการทำดาต้าไมนิ่งโดยนำหลักการและรูปแบบของต้นไม้ช่วยตัดสินใจโดยใช้อัลกอริทึม SLIQ

1.4 เครื่องมือที่ใช้ในการพัฒนาระบบ

1. เครื่องมือที่ใช้ในการเขียน โปรแกรม Visual Basic 6.0
2. ระบบฐานข้อมูล Microsoft Access 97
3. คอมพิวเตอร์รุ่น Pentium III, 2 CPU 1.6 GHz, RAM 1 GB.
4. ระบบปฏิบัติการ Windows XP V.2002

1.5 ขั้นตอนและวิธีการดำเนินงาน

1. ศึกษาแนวคิดและขั้นตอนของการทำดาต้าไมนิ่ง
2. ศึกษาเทคนิคจัดหมวดหมู่ของข้อมูล (Data Classification)
3. ศึกษาหลักการและรูปแบบของต้นไม้ช่วยตัดสินใจโดยใช้อัลกอริทึม SLIQ
4. รวบรวมข้อมูลและเตรียมข้อมูลที่จะนำมาใช้ในการพัฒนาระบบตามหลักการของการทำดาต้าไมนิ่ง
5. พัฒนาระบบ
6. ทดสอบ ตรวจสอบและแก้ไขความถูกต้องของระบบ
7. จัดทำเอกสารประกอบและคู่มือ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. ผู้พัฒนาระบบมีความรู้ และสามารถนำเทคนิคของการทำดาต้าไมนิ่งที่ได้ศึกษาไปประยุกต์ใช้ในการวิเคราะห์ข้อมูล
2. เพื่อช่วยวิเคราะห์ข้อมูลและลดระยะเวลาในการปฏิบัติงานของผู้ใช้
3. เพื่อนำข้อมูลที่มีอยู่มาใช้ให้เกิดประโยชน์สูงสุดและเพิ่มประสิทธิภาพของการทำงานขององค์กร



บทที่ 2

การบริหารลูกค้าสัมพันธ์และการประเมินพฤติกรรมด้านสินเชื่อ

2.1 การบริหารลูกค้าสัมพันธ์ (Customer Relationship Management: CRM)

การบริหารลูกค้าสัมพันธ์คือการสร้างความสัมพันธ์กับลูกค้า โดยการใช้เทคโนโลยีและการใช้บุคลากรอย่างมีประสิทธิภาพ เพื่อเพิ่มระดับการให้บริการแก่ลูกค้า สร้างความเข้าใจในความต้องการของลูกค้า รวมทั้งตอบสนองต่อความต้องการของลูกค้าทั้งในด้านผลิตภัณฑ์ และการบริการ

เป้าหมายของการบริหารลูกค้าสัมพันธ์นั้น ไม่ได้เน้นเพียงแค่การบริการลูกค้าเท่านั้น แต่ยังรวมถึงการเก็บข้อมูลพฤติกรรมในการใช้จ่ายและความต้องการของลูกค้า จากนั้นจะนำข้อมูลเหล่านั้นมาวิเคราะห์และใช้ให้เกิดประโยชน์ในการพัฒนาผลิตภัณฑ์ หรือการบริการ รวมไปถึงนโยบายในการจัดการ ซึ่งเป้าหมายสุดท้ายของการพัฒนาการบริหารลูกค้าสัมพันธ์ก็คือ การเปลี่ยนจากผู้บริโภคไปสู่การเป็นลูกค้าตลอดไป (Customer loyalty)

2.1.1. ประโยชน์ของการบริหารลูกค้าสัมพันธ์

1. มีรายละเอียดข้อมูลของลูกค้าในด้านต่างๆ ได้แก่ Customer Profile, Customer Behavior
2. วางแผนทางการตลาดและการขายอย่างเหมาะสม
3. ใช้กลยุทธ์ในการตลาด และการขายได้อย่างรวดเร็วและมีประสิทธิภาพตรงความต้องการของลูกค้า
4. เพิ่มและรักษาส่วนแบ่งตลาดของธุรกิจ
5. ลดการทำงานที่ซ้ำซ้อน ลดค่าใช้จ่ายและเพิ่มประสิทธิภาพของการทำงาน เพิ่มโอกาสในการแข่งขันก่อให้เกิดภาพพจน์ที่ดีต่อองค์กรเพราะลูกค้าที่ดีในปัจจุบัน อาจเป็นลูกค้าที่ไม่ดีหรือไม่สร้างกำไรต่อองค์กรในอนาคตก็ได้ หรือลูกค้าที่ยังไม่ทำอะไรต่อเราในปัจจุบัน อาจเป็นลูกค้าที่ดีของเราในอนาคตก็ได้

2.1.2. ประโยชน์ของระบบการบริหารลูกค้าสัมพันธ์ที่ดี

1. เก็บข้อมูลลูกค้า และสามารถวิเคราะห์ข้อมูลโดยทันที ซึ่งจะเป็นประโยชน์อย่างยิ่งต่อการออกแบบสินค้า หรือบริการ รวมทั้งการวางแผนกลยุทธ์ทางการตลาด
2. สามารถให้บริการที่ยอดเยี่ยมแก่ลูกค้า เช่น การปรับแต่งการให้บริการที่เหมาะสมกับลูกค้า การตอบข้อซักถาม หรือส่งสินค้าและบริการถึงมือลูกค้าได้อย่างรวดเร็ว เป็นต้น
3. สามารถระบุกลุ่มลูกค้าที่สร้างผลกำไรสูงสุดให้แก่องค์กรได้แม่นยำ และมุ่งความสนใจ

เอกสารนี้ไปยังลูกค้ากลุ่มนี้ ทำให้สามารถหลีกเลี่ยงธุรกิจที่ไม่สร้างผลกำไร
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้คัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4. ลดค่าใช้จ่ายในการปฏิสัมพันธ์กับลูกค้า การวิเคราะห์ข้อมูลลูกค้า หรือการสร้างรายงาน

2.2 การทำคาด้าไมนิ่งเพื่อการบริหารลูกค้าสัมพันธ์มีหลายประเภทดังนี้

2.2.1. การจัดแคมเปญ

การจัดแคมเปญนั้นเพื่อนำเสนอรายการส่งเสริมการขายให้ตรงกับพฤติกรรมการใช้งานหรือความต้องการของลูกค้า ซึ่งการทำคาด้าไมนิ่งเป็นวิธีการหนึ่งที่น่ามาใช้ในการวิเคราะห์เพื่อให้ได้แคมเปญที่นำเสนอเหมาะกับพฤติกรรมหรือความต้องการของลูกค้าอย่างแท้จริง และเป็นส่วนช่วยเพิ่มยอดขาย

2.2.2. การแบ่งกลุ่มของลูกค้า

การแบ่งกลุ่มของลูกค้านั้นเพื่อประโยชน์ในการปรับปรุงสินค้า การบริการ และข้อความทางการตลาดที่จะใช้สื่อสารไปยังลูกค้าแต่ละกลุ่ม การแบ่งกลุ่มของลูกค้าแบบเก่านั้นจะใช้การแบ่งกลุ่มจากการสำรวจประชากรหรือการวิจัยตลาด เช่น อายุ โสด หรือ กลุ่มที่มีความซื่อสัตย์ต่อแบรนด์สูง เป็นต้น ซึ่งปัญหาที่พบในการแบ่งกลุ่มแบบจากการวิจัยตลาดนั้นยากที่จะนำผลสำรวจมาประยุกต์ใช้กับกลุ่มลูกค้าที่ไม่ได้เป็นส่วนหนึ่งในการสำรวจดังกล่าว ส่วนปัญหาจากการแบ่งกลุ่มลูกค้าจากการสำรวจประชากรนั้นคือเราไม่อาจจะสันนิษฐานได้ว่าแท้จริงแล้วลูกค้าแต่ละรายจะมีรสนิยมหรือความพึงพอใจในสินค้าหรือบริการเหมือนลูกค้าคนอื่นๆ ในกลุ่มนั้น ด้วยวิธีการทำคาด้าไมนิ่งจะสามารถที่จะแบ่งพฤติกรรมของลูกค้าในแต่ละกลุ่มได้เป็นอย่างดี

2.2.3. การลดความเสี่ยงเครดิต และการติดตามหนี้

การเรียนรู้พฤติกรรมของลูกค้าที่ไม่ดี และการสังเกตกลุ่มลูกค้าที่ดีที่อาจจะกลายเป็นลูกค้าที่ไม่ดีนั้น เป็นสิ่งที่สำคัญเช่นเดียวกันกับการรักษากลุ่มลูกค้าที่ดีที่มีอยู่ ซึ่งวิธีการลดความเสี่ยงเครดิตนั้นมักจะทำด้วยวิธีการตรวจสอบเรื่องเครดิตของลูกค้าเบื้องต้น ซึ่งเป็นไปได้ว่ารูปแบบของความเสียหายของลูกค้าแต่ละรายอาจจะเปลี่ยนไปก็ได้ และหากลูกค้าไม่ชำระค่าสินค้าหรือบริการ การทำคาด้าไมนิ่งก็สามารถที่จะนำมาใช้ประโยชน์ในการพยากรณ์จำนวนเงินที่จะสามารถติดตามหนี้ และในบางกรณีนั้นยังสามารถช่วยวางกลยุทธ์ที่จะใช้ติดตามหนี้อีกด้วย

2.2.4. การกำหนดค่าของลูกค้า

การกำหนดค่าของลูกค้าก่อนข้างจะซับซ้อน แม้ว่าจะใช้วิธีการทำคาด้าไมนิ่ง เป็นเพราะว่าบางครั้งจะกำหนดจากค่าจำกัดความทางการเงิน เช่น อาจะกำหนดเป็นสมการดังนี้

ค่าของลูกค้า = รายได้ทั้งหมดที่ได้จากลูกค้า - ค่าใช้จ่ายทั้งหมดในการบำรุงรักษาลูกค้า

เอกสาร์เป็นเอกสารที่สงวนไว้สำหรับงานวิชาการเรียนเท่านั้น ไม่อนุญาตให้ทำในเชิงพาณิชย์
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้คัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หรืออาจจะกำหนดแบบง่ายๆว่าเป็นค่าใช้จ่ายของลูกค้าในแต่ละวัน แต่ละเดือน เป็นต้น

2.2.5. การเสนอขายแบบต่อเนื่อง (Cross-selling) และการเสนอขายแบบเพิ่มเติม (Up-selling)

ต้นทุนในการหาลูกค้าใหม่เป็นค่าใช้จ่ายที่สูงสำหรับธุรกิจ ดังนั้นการบริหารลูกค้าสัมพันธ์ส่วนมากจะให้ความสนใจเกี่ยวกับเรื่องการเพิ่มยอดขายได้และเพิ่มจำนวนลูกค้าด้วยการทำการเสนอขายแบบต่อเนื่องและแบบต่อ ซึ่งการทำคาด้าไม้นั้นจะใช้หาว่าเราควรจะนำเสนออะไรให้กับลูกค้าคนใด และเมื่อไหร่ที่ควรจะเสนอแก่ลูกค้าเป็นต้น

1. การเสนอขายแบบต่อเนื่องเป็นกระบวนการเสนอ การแนะนำ หรือเชิญชวน รวมถึงการจูงใจให้ลูกค้าผู้ตัดสินใจซื้อผลิตภัณฑ์ของเราไปแล้ว พิจารณาซื้อผลิตภัณฑ์อื่นๆ เพิ่มอีก ไม่ว่าจะเป็นสินค้าหรือบริการของเราเองที่มีอยู่ ของแผนกอื่น ของบริษัทในเครือ หรือแม้แต่ของพันธมิตรธุรกิจของเราก็ตาม

เป้าหมายของการทำการเสนอขายแบบต่อเนื่องก็คือการเพิ่มโอกาสที่จะขายบริการในส่วนอื่นๆ แก่ลูกค้า และโดยมาก การตอบสนองของลูกค้ามักเป็นไปในด้านบวก เช่น เมื่อลูกค้าทุกรายชำระเงินค่าสินค้าหรือบริการนั้น พนักงานร้านสะดวกซื้อก็จะต้องมีประโยศ่งท้ายว่า “รับขนมจีบ ซาลาเปา เพิ่มด้วยมั๊ยคะ?” เป็นต้น

การเสนอขายแบบต่อเนื่องอาจใช้ร่วมกับเครื่องมืออื่นๆ เช่น Direct-mail, Telemarketing, Email Newsletter, Internet Web Site เป็นต้น

2. การเสนอขายเพิ่มเติมเป็นกระบวนการที่ออกแบบสำหรับการเพิ่มเงิน ณ จุดขาย โดยการเสนอการอัพเกรดสินค้า หรืออุปกรณ์ที่เกี่ยวข้อง ณ จุดที่ลูกค้าซื้อครั้งแรก หรือนำเสนอสินค้าราคาพิเศษถ้าสั่งซื้อสินค้าเดี๋ยวนี

ธุรกิจบางแห่งหลีกเลี่ยงวิธีการเสนอขายเพิ่มเติมเพราะเกรงว่าจะเป็นการสร้างความรำคาญ และลูกค้าอาจจะยกเลิกการซื้อสินค้าทั้งหมด แต่จากประสบการณ์พบว่าลูกค้าจะพอใจเพราะเป็นการเอาใจใส่ในการนำเสนอผลประโยชน์ที่เพิ่มขึ้นจากสิ่งที่เขาไม่ทราบมาก่อน สิ่งที่สามารถนำเสนอเมื่อลูกค้าตัดสินใจซื้อครั้งแรก ได้แก่

- ส่วนลดของสินค้าและบริการที่เกี่ยวข้อง
- ส่วนลดจากปริมาณ ถ้ามีการซื้อเพิ่มขึ้น
- ราคาพิเศษจากแพ็คเกจที่รวมทั้งสินค้าและบริการที่ซื้อเดิม

จากการสำรวจธุรกิจจำนวนมากพบว่าอัตราการยอมรับข้อเสนอของการเสนอขายเพิ่มเติมอยู่ระหว่างร้อยละ 40-60 ซึ่งค่อนข้างสูงจากการใช้ความพยายามไม่สูงมากนัก นอกจากนี้กำไรของการเสนอขายเพิ่มเติมจะสูงกว่าการขายสินค้าและบริการครั้งแรกเพราะไม่มีค่าใช้จ่ายเพิ่มขึ้น

ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2.6. กลุ่มลูกค้าที่มีแนวโน้มที่จะเลิกใช้สินค้าหรือบริการ และการรักษาลูกค้า (Churn and Retention)

การวิเคราะห์กลุ่มลูกค้าที่มีโอกาสที่จะเลิกใช้สินค้าหรือบริการนั้นมักจะเป็นหัวข้อหลักในโปรแกรมประยุกต์ของร้านค้าปลีกที่มีในท้องตลาด ซึ่งการทำค้าปลีกนั้นจะสามารถพยากรณ์ได้ว่าลูกค้ารายใดที่มีแนวโน้มที่จะเลิกใช้สินค้าหรือบริการ หรือพยากรณ์ว่าลูกค้าแต่ละรายนั้นจะอยู่กับเรานานแค่ไหนซึ่งการรักษาลูกค้านั้นเราจะต้องนำเสนอสินค้าเพื่อจูงใจลูกค้ากลุ่มนี้

2.3 การประเมินพฤติกรรมด้านสินเชื่อ

2.3.1. การให้คะแนนความสามารถในการชำระหนี้ (Credit Scoring)

การประเมินพฤติกรรมด้านสินเชื่อโดยใช้วิธีการให้คะแนนความสามารถในการชำระหนี้ (Credit Scoring) เป็นระบบที่ธนาคาร สถาบันการเงิน ผู้ให้กู้ยืม ผู้รับประกัน เจ้าของที่ดิน นายจ้าง และบริษัทสาธารณูปโภค มักใช้ในการประเมินพฤติกรรมของลูกค้าโดยพิจารณาว่าลูกค้าแต่ละรายมีความน่าเชื่อถือหรือไม่ ข้อพิจารณานี้จะมีความสำคัญยิ่งขึ้นเมื่อลูกค้าต้องการสมัครขอสินเชื่อแบบไม่มีหลักทรัพย์ค้ำประกัน เช่นวงเงินสินเชื่อ หรือบัตรเครดิตผู้ให้สินเชื่อจะรวบรวมข้อมูลเกี่ยวกับตัวลูกค้าและประวัติด้านสินเชื่อจากใบสมัครขอสินเชื่อ ซึ่งรายงานข้อมูลเครดิตของลูกค้าแต่ละรายคะแนนความสามารถในการชำระหนี้จะขึ้นอยู่กับประวัติการจ่ายเงินตามใบแจ้งหนี้ จำนวนเงินประเภทบัญชีที่ลูกค้าถือครองอยู่ในปัจจุบัน การจ่ายชำระคืนล่าช้า การดำเนินการเรียกเก็บหนี้ (ในกรณีที่ลูกค้าได้สมัครขอสินเชื่อใหม่เมื่อไม่นานมานี้) หนี้สินคงค้างและระยะเวลาที่ลูกค้าได้ถือครองบัญชีที่ถืออยู่ในปัจจุบัน

ด้วยโปรแกรมทางสถิติทำให้สามารถเปรียบเทียบข้อมูลของลูกค้าและพฤติกรรมการจ่ายคืนสินเชื่อของลูกค้าคนอื่นๆที่มีประวัติใกล้เคียงกัน ระบบดังกล่าวจะให้คะแนนปัจจัยแต่ละอย่างที่น่าสงสัยว่าใครมีแนวโน้มจะจ่ายคืนหนี้สิน คะแนนรวมความสามารถในการชำระหนี้ที่ได้จะช่วยในการคาดการณ์ว่าลูกค้ามีความน่าเชื่อถือมากน้อยแค่ไหน นั่นคือ มีแนวโน้มมากน้อยเพียงใดที่ลูกค้าจะจ่ายคืนเมื่อถึงกำหนด เป็นต้น ปัจจุบันลูกค้าเองสามารถที่จะตรวจสอบคะแนนความสามารถในการชำระหนี้ของตัวเองได้จากบริษัทข้อมูลเครดิต

2.3.2. ข้อมูลที่ใช้พิจารณาคะแนนความสามารถในการชำระหนี้มีดังนี้

1. ประวัติของลูกค้า (Customer History) โดยดูประวัติของลูกค้าว่า ลูกค้ามีความรับผิดชอบมากน้อยแค่ไหนในการจ่ายเงินตามใบแจ้งหนี้เมื่อถึงกำหนดชำระ
2. ความสามารถในการจ่ายเงินหรือหนี้สิน (Capacity) โดยพิจารณาความสามารถจ่ายคืนเงินกู้

เอกสารนี้เป็นเอกสารของบริษัท ข้อมูลนี้เป็นข้อมูลเบื้องต้น ไม่สามารถนำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. หลักทรัพย์ค้ำประกัน (Collateral) โดยพิจารณาหลักทรัพย์ประกันการกู้ยืมในกรณีที่ลูกค้าไม่จ่ายคืนเงินกู้ ตัวอย่างเช่น บ้านที่ใช้เป็นหลักทรัพย์ค้ำประกันเงินกู้จำนอง
- โดยทั่วไปแล้วการพิจารณาดังกล่าวจะมีการปรับคะแนนความสามารถในการชำระหนี้ของลูกค้าซึ่งจะขึ้นอยู่กับประวัติของลูกค้าและความสามารถในการจ่ายเงินหรือหนี้สินที่ผ่านมา



บทที่ 3

ดาต้าไมนิ่ง

3.1 ความหมายของดาต้าไมนิ่ง

ดาต้าไมนิ่งเป็นการกระบวนการค้นหาความสัมพันธ์ การเปลี่ยนแปลง ความผิดปกติและรูปแบบทั้งหมดที่มีอยู่ในฐานข้อมูลซึ่งสัมพันธ์กัน แต่ความสัมพันธ์และรูปแบบเหล่านั้นได้ถูกซ่อนไว้ภายในฐานข้อมูลที่มีปริมาณข้อมูลขนาดใหญ่ ดาต้าไมนิ่งจะทำการสำรวจและวิเคราะห์ข้อมูลให้อยู่ในรูปแบบที่เต็มไปด้วยความหมายและอยู่ในรูปของกฎ โดยความสัมพันธ์เหล่านี้แสดงให้เห็นถึงความรู้ต่างๆ (Knowledge) ที่มีประโยชน์ ดาต้าไมนิ่งจึงเป็นที่รู้จักและนำมาใช้ประยุกต์ใช้อย่างแพร่หลาย เนื่องจากดาต้าไมนิ่งสามารถดึงความรู้ออกมาจากข้อมูลจำนวนมากที่ถูกเก็บสะสมและซ่อนไว้ในฐานข้อมูลขององค์กร เราอาจจะเรียกดาต้าไมนิ่งอีกอย่างหนึ่งว่า “Knowledge Discovery in Databases” (KDD) (Piatesky-Shapiro and Frawley, 2001)

3.2 การทำดาต้าไมนิ่งประกอบด้วย 6 ขั้นตอนหลักๆ ดังนี้

3.2.1 Problems formulation

การกำหนดวัตถุประสงค์ทางธุรกิจ คือจะต้องเข้าใจปัญหาและความต้องการทางธุรกิจ การกำหนดวัตถุประสงค์ เป็นการระบุขอบเขตของข้อมูลที่จะนำมาทำการวิเคราะห์ทางธุรกิจนั้นจะเป็นส่วนที่กำหนดว่าเมื่อไหร่ที่จะใช้ดาต้าไมนิ่งในการแก้ปัญหาซึ่งในส่วนนี้จะประกอบด้วยการวิเคราะห์ทางธุรกิจและการวิเคราะห์เบื้องต้นว่าเรามีข้อมูลโดยอยู่บ้างและต้องการอะไรจากข้อมูล ซึ่งขั้นตอนนี้จะสามารถมองถึงอัลกอริทึมและฐานข้อมูลที่สัมพันธ์กับวัตถุประสงค์ทางธุรกิจได้

การใช้งานดาต้าไมนิ่งให้ได้ประโยชน์สูงสุดจำเป็นต้องมีการกำหนดวัตถุประสงค์ที่ชัดเจน เช่น ต้องการเพิ่มยอดขายทางจดหมาย ขึ้นอยู่กับการระบุเป้าหมายว่าจะเพิ่มอัตราการตอบรับหรือเพิ่มมูลค่าการตอบรับ ซึ่งจำเป็นที่จะต้องสร้างโมเดลที่แตกต่างกัน วัตถุประสงค์ที่กำหนดขึ้นมาจะต้องมีการระบุวิธีการในการวัดผลลัพธ์ที่ได้จากโครงการ รวมถึงต้นทุนที่สมเหตุสมผลด้วย

3.2.2 Data Selection and Preparation

เป็นหัวใจของขั้นตอนการทำดาต้าไมนิ่ง และเป็นช่วงที่ใช้เวลามากที่สุด โดยปกติแล้วต้องการเวลาประมาณ 60% ของเวลาทั้งหมด ในการเตรียมข้อมูลในขั้นตอนนี้สามารถแบ่ง ออกได้เป็น 5 ขั้นตอนย่อยดังต่อไปนี้

1. การเลือกข้อมูล (Data Selection) จุดประสงค์ของการเลือกข้อมูล คือ การระบุแหล่งของข้อมูลที่มีและทำการดึงเอาข้อมูลออกมาใช้สำหรับการวิเคราะห์เบื้องต้นในการ เตรียมตัวสำหรับการที่จะทำการไมนิ่งในขั้นต่อไป การเลือกข้อมูลนั้นจะแตกต่างกันไปตามวัตถุประสงค์ของแต่ละ

ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ธุรกิจที่ได้กำหนดไว้ตั้งแต่ต้นและการเลือกข้อมูลก็ยังถูกกำหนดโดยลักษณะงานที่จะถูกนำมาใช้อีกด้วย

ตัวแปรที่ถูกเลือกมาแต่ละตัวนั้นจะต้องถูกทำความเข้าใจว่าตัวแปรแต่ละตัวหมายความว่าอะไร ประกอบด้วยอะไร ไม่เพียงแต่คำจำกัดความทางธุรกิจเท่านั้น แต่จะต้องมีคำอธิบายอย่างชัดเจนเกี่ยวกับชนิดของข้อมูล ค่าที่เป็นไปได้ แหล่งกำเนิดของข้อมูล รูปแบบของข้อมูล และลักษณะอื่นๆ ซึ่งจะมีตัวแปร 2 ชนิดคือ

- ตัวแปรแบบ Categorical ประกอบด้วย

1. Nominal Variable กล่าวถึงชนิดนี้ของ Object ที่มันอ้างถึงแต่ไม่มีลำดับในค่าที่เป็นไปได้ (Possible Value) ตัวอย่างเช่น สถานะการแต่งงาน (โสด, แต่งงาน, หย่า, ไม่ทราบ) เพศ (ชาย, หญิง) ระดับการศึกษา (ปริญญาโท, ปริญญาตรี, ม. ปลาย, ปวช)
2. Ordinal Variable มีลำดับสำหรับค่าที่เป็นไปได้ ตัวอย่างเช่น ลำดับของลูกค้า (ดี, ปานกลาง, ไม่ดี)

- ตัวแปรแบบ Quantitative ซึ่งมีการวัดความแตกต่างระหว่างค่าที่เป็นไปได้ ประกอบด้วย

1. Continuous (ค่าที่ต่อเนื่อง) เช่น รายได้เฉลี่ย จำนวนครั้งที่ซื้อ รายได้
2. Discrete (ค่าเป็นจำนวนเต็ม) เช่น จำนวนพนักงาน เวลาปี (เดือน, ฤดู, ไตรมาส)

ตัวแปรของข้อมูลมีหลายตัวมาก แต่ตัวแปรที่ถูกเลือกสำหรับการทำค้ำไ่มนึ่งนั้นถูกเรียกว่า “Active Variable” เพราะว่ามันจะถูกใช้สร้างความแตกต่างของกลุ่มย่อยต่างๆ และสามารถถูกนำมาพยากรณ์ผลได้ เมื่อคุณทำการเลือกข้อมูลจะต้องพิจารณาอายุของข้อมูลด้วย เพราะว่าการเปลี่ยนแปลงตลอดเวลาซึ่งจะทำให้ประสิทธิภาพของการทำไ่มนึ่งลดลง ตัวอย่างเช่น ธรรมเนียมการใช้ชีวิต การเปลี่ยนงาน เป็นต้น

2. การกลั่นกรองข้อมูล (Data Preprocessing) จุดประสงค์ก็เพื่อให้มั่นใจว่าคุณภาพของข้อมูลที่ถูกเลือกนั้นเหมาะสม ข้อมูลที่สมบูรณ์เป็นเครื่องประกันว่าการทำค้ำไ่มนึ่งจะสำเร็จ ในขั้นตอนนี้เป็นขั้นตอนที่มีปัญหามากกว่าในขั้นตอนของการเตรียมข้อมูล เพราะข้อมูลส่วนใหญ่ที่มีในองค์กร ไม่ได้ถูกเตรียมมาเพื่องานค้ำไ่มนึ่ง โดยเฉพาะ ข้อมูลจะถูกนำมาจากแหล่งต่างๆ ถูกจัดเก็บไม่ดี ข้อมูลที่ถูกนำมาจากภายนอก แล้วนำมาเพื่อให้เข้ากับข้อมูลภายในที่มีอยู่ ปัญหาหลักของข้อมูลคือ คุณภาพและ Data Integrity

ในขั้นตอนนี้ก่อนอื่นจะต้องทำการทบทวนโครงสร้างของข้อมูลใหม่และวัดคุณภาพของมัน โดยวิธีทางสถิติ หรือสูตรตัวอย่างเครื่องมือที่ใช้ในการการกลั่นกรองข้อมูลมีดังต่อไปนี้

- ค่าตัวแปรเป็นแบบ Categorical การแบ่งความถี่ของค่าจะเป็นวิธีที่ทำให้เกิดความเข้าใจใน Data Content เครื่องมือทางด้านกราฟฟิคจะเป็นตัวช่วยให้เห็นและกำหนดค่าที่หายไปได้

• ตัวแปรแบบ Quantitative ตัวแปรประเภทนี้มักมีการใช้การวัด ตัวอย่างเช่น ค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย ค่ากลาง ค่ามัธยฐาน และค่าอื่นๆ ทางสถิติ เมื่อนำค่าพวกนี้มาเข้าสู่ตรรกานวนก็จะบอกถึงค่าที่ไม่สมบูรณ์ หรือค่าที่มีปัญหา

ระหว่างการทำขั้นตอนการกลั่นกรองข้อมูลจะมีปัญหาบ่อยๆ ที่มักพบได้เรียกว่า “Noisy Data” คือตัวแปรตัวหนึ่งหรือมากกว่ามีค่าซึ่งเกินกว่าค่าที่เราคาดไว้ ซึ่งอาจจะหมายถึงแง่ดีหรือแง่ร้ายก็ได้ ในแง่ดีก็คือ มันจะแสดงอย่างชัดเจนถึงโอกาสซึ่งเรากำลังมองหาอยู่ ในแง่ร้าย คือมันอาจจะเป็นข้อมูลที่ไม่สมบูรณ์ สาเหตุที่เกิดขึ้นได้อาจจะมาจากความเลินเล่อของมนุษย์ ตัวอย่างเช่น การใส่อายุให้คนเป็น 300 ปี หรือใส่ค่าของรายได้เป็นค่าลบ ค่าเหล่านี้ควรจะถูกแก้ไข หรือเอาออกจากการวิเคราะห์ และควรมีขั้นตอนการตรวจสอบข้อมูลก่อนนำมาใช้

ค่าที่หายไป Missing Value คือค่าที่ไม่ได้แสดงในข้อมูลที่เราได้เลือกไว้แล้ว หรือค่าที่ไม่สมบูรณ์ที่เราได้ลบออกไประหว่าง การทำ Noise Detection ค่าอาจจะหายไปเพราะเกิดจากความเลินเล่อของมนุษย์ เพราะว่าไม่มีข้อมูลนั้นระหว่างการทำ Input Data การจัดการกับค่าที่หายไป นั้นสามารถจัดการได้ด้วยเทคนิคที่ต่างๆ กัน

3. การสำรวจและตรวจสอบข้อมูล (Data Cleaning and exploration) เมื่อทำการเก็บข้อมูลเรียบร้อยแล้ว ขั้นตอนต่อไปที่ควรกระทำก็คือการตรวจสอบข้อมูล เหตุที่ต้องทำการตรวจสอบข้อมูลมี 2 ข้อคือ

1. นักวิเคราะห์ควรมีความคุ้นเคยกับตัวข้อมูล ไม่ใช่รู้แค่ชื่อของ Attribute และความหมายของมันเท่านั้น แต่ต้องรู้ถึงเนื้อหา (Content) หรือความมุ่งหมายที่แท้จริงของข้อมูลด้วย

2. อาจมีความผิดพลาดของการเก็บสะสมข้อมูลเกิดขึ้นในขณะที่ทำการรวบรวมข้อมูลจากฐานข้อมูลหลายๆ แหล่งเข้ามาเป็นหนึ่งเดียวเพื่อใช้ในการวิเคราะห์ ซึ่งนักวิเคราะห์ที่ดีจะต้องทำการตรวจสอบข้อมูลเหล่านี้ให้ถูกต้อง ตัวอย่างของความผิดพลาดที่เกิดขึ้น ได้แก่ ความผิดพลาดในการเก็บข้อมูลจาก Attribute ที่ไม่ต้องการ ซึ่งเกิดจากความสับสนในการตั้งชื่อ Attribute นั้น (Mislabeling of field) ตัวอย่างเช่น เราต้องการเก็บค่าของระดับการศึกษาของผู้สมัครเข้าศึกษาต่อ ซึ่งในความเป็นจริงถูกเก็บไว้ใน Attribute ที่ชื่อ “LEVEL_EDU” แต่ในฐานข้อมูลนั้นบังเอิญมี Attribute อีกตัวหนึ่งชื่อ “EDUCATION” ซึ่งเก็บระดับการศึกษาที่ผู้สมัครต้องการเข้าศึกษาต่อ ซึ่งถ้าเราไม่ได้ตรวจสอบความสัมพันธ์และความมุ่งหมายที่แท้จริงของแต่ละ Attribute แล้ว ก็อาจเกิดความสับสน โดยเก็บข้อมูลของ Attribute “EDUCATION” ไปแทนก็ได้ ซึ่งเมื่อนำข้อมูลที่ไปทำคาด้าไม่นิ่งผลลัพธ์ที่ได้ ก็จะผิดพลาดด้วย

4. การแปลงข้อมูล (Data Transformation) ระหว่างขั้นตอนของการแปลงข้อมูลข้อมูลที่ได้กลั่นกรองแล้วจะถูกแปลงให้เป็นรูปแบบของข้อมูลที่จะถูก วิเคราะห์ รูปแบบของข้อมูลที่ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ตัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

พร้อมจะถูกระบุว่า คือรูปแบบของข้อมูลที่ไม่มีความชัดเจน ถูกจัดระเบียบมาอย่างเรียบร้อย กลับกันการมองจากแหล่งข้อมูลภายนอกและแหล่งข้อมูลภายใน

ขั้นตอนนี้เป็นขั้นตอนที่สำคัญมากเนื่องจากความถูกต้อง และสมบูรณ์ของผลลัพธ์สุดท้ายซึ่งขึ้นอยู่กับว่า นักวิเคราะห์ ข้อมูลนั้นตัดสินใจกำหนดโครงสร้างและเสนอลักษณะของ Input อย่างไร ตัวอย่างเช่น หลักการรูปแบบของข้อมูลถูกกำหนดแล้ว ข้อมูลที่ถูกกลั่นกรองจะเหมาะสมกับรูปแบบเฉพาะสำหรับแต่ละกรรมวิธีของค้ำไ่มันที่จะถูกใช้ การแปลงข้อมูลยังรวมไปถึงการทำ Data Recording และ Data Format Conversion เช่นการแปลงวันที่ เป็นต้น

ทางสถิติการทำการแปลงข้อมูลยังมีเทคนิคของ Data Reduction จุดประสงค์เพื่อที่จะลดตัวแปรสำหรับการทำการ Process โดยการนำเอาตัวแปรตั้งแต่ 2 ตัวขึ้นไปมารวมกันแล้ว ทำการ Process ข้อดีก็คือลดจำนวนของตัวแปรลง และยังสามารถจัดการได้ง่ายขึ้น

อีกเทคนิคเรียกว่า Discretization โดยการแปลงตัวแปรแบบ Quantitative ให้เป็นแบบ Categorical โดยการแบ่ง ค่าของตัวแปรที่จะเป็น Input ให้เป็นช่วงๆ เช่นการแปลงเงินเดือน อายุ

อีกเทคนิคเรียกว่า One of N โดยการแปลงตัวแปรแบบ Categorical ให้เป็น Numeric ตัวอย่างเช่น ชนิดของรถ Ford, Lincoln, Nissan ให้เป็น 100, 010, 001 ปกติแบบนี้มักจะเป็น Input ของพวก Neural Network

5. การปรับแต่งข้อมูล (Data Engineering) ขั้นตอนก่อนหน้านี้นี้เป็นขั้นตอนของการสร้าง และการตรวจสอบความถูกต้องของข้อมูลที่จะนำมาใช้ แต่ในขั้นตอนนี้ที่เราต้องทำ คือการปรับแต่งฐานข้อมูล ซึ่งในขั้นตอนนี้จะมีปัญหาหลักๆ อยู่ 3 ข้อคือ

1. ฐานข้อมูลที่ได้ อาจมี Attributes จำนวนมากที่สามารถใช้ประโยชน์ได้แต่ถูกละเลย การเลือกกลุ่มของ Attributes ที่จะใช้เป็นปัญหาที่สำคัญปัญหาหนึ่ง
2. ฐานข้อมูลที่ได้ อาจมีจำนวนระเบียน (record) มากเกินไปกว่าที่จะสามารถทำการวิเคราะห์ให้เสร็จลงได้ในเวลาที่เหมาะสม ซึ่งในกรณีนี้เราต้องทำการสุ่มข้อมูลตัวอย่างขึ้นมาใช้แทน
3. ข้อมูลบางอย่างอาจใช้ให้เกิดประโยชน์ได้ โดยการนำเสนอในรูปแบบของการวิเคราะห์แบบเฉพาะเจาะจง การทำ Data engineering นั้นจะมีการทำซ้ำเข้ามาหลายๆ ครั้งเพื่อทดสอบการใช้ Attribute ที่แตกต่าง ขนาดของกลุ่มตัวอย่างที่ต่างกัน เช่น เราจะพยากรณ์อนาคตเมื่อเวลาผ่านไป 1, 2, 3 หรือ 4 เดือน เราอาจพยากรณ์ได้โดยใช้เพียง Attribute เป็นตัวพยากรณ์ หรืออาจใช้ข้อมูลทุกอย่างที่เราเป็นตัวพยากรณ์ก็ได้ เป็นต้น

3.2.3 Visualization

เป็นการนำเสนอข้อมูลในรูปแบบกราฟฟิก การนำเสนออาจจะสามารถทำได้มากกว่า 2 มิติ ซึ่งจะสร้างความละเอียดของการนำเสนอและสร้างความเข้าใจให้มากขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2.4 Analysis

หลังจากเลือกอัลกอริทึมที่เหมาะสมกับลักษณะของปัญหาแล้ว เราก็จะนำอัลกอริทึมนั้นมาทำการวิเคราะห์ข้อมูลในฐานข้อมูลที่เตรียมไว้ ขั้นตอนนี้จะถูกเรียกว่า คาด้าไมนิ่ง (Data Mining) ผลลัพธ์ที่ได้จากขั้นตอนนี้จะป็นรูปแบบของความสัมพันธ์ของข้อมูลที่จะนำมาใช้ ในการพยากรณ์ (Prediction) หรือวิเคราะห์ต่อไป

การนำข้อมูลที่จัดเตรียมไว้นามาวิเคราะห์ตามกระบวนการทำคาด้าไมนิ่งนั้นมี 4 แบบคือ

1. Database Segmentation เป็นกระบวนการแบ่งฐานข้อมูล ออกเป็นกลุ่มเพื่อให้่ายต่อการวิเคราะห์ เช่นการแบ่งลูกค้าออกตามอายุ เพศ รายได้ เป็นต้น โดยหนังสือบางเล่มอาจใช้คำว่า Clustering แทนคำว่า Segmentation

วัตถุประสงค์ของการแบ่งฐานข้อมูลเพื่อแบ่งส่วนจำนวนคุณลักษณะและถูกพิจารณาเป็นส่วนที่เหมือนกันของข้อมูลที่สัมพันธ์กันในแต่ละฐานข้อมูล หรือก่อนที่จะทำการสืบค้นข้อมูล เช่น การทำ Predictive Modeling

2. Predictive Modeling เป็นการสร้างแบบจำลองพยากรณ์ แบ่งเป็น 2 ลักษณะ คือ

- Classification เป็นการจัดหมวดหมู่ให้กับแต่ละข้อมูลในฐานข้อมูล โดยมีการระบุค่า หรือลักษณะที่เป็นไปได้ของข้อมูล ภายในแต่ละกลุ่ม เช่น การจัดกลุ่มของผู้ป่วยตามผลของการใช้ยา รักษา เพื่อระบุรูปแบบการรักษาให้กับผู้ป่วยใหม่ที่เข้ารับการรักษา เป็นต้น
- Value Prediction เป็นการพยากรณ์ค่าที่เป็นไปได้ หรือการกระจายของค่าที่เป็นไปได้ของตัวแปรใดๆ ในกลุ่มข้อมูลการพยากรณ์ค่าที่เป็นตัวเลข เช่น การพยากรณ์ภาษีที่จะเก็บได้ในปีงบประมาณ เป็นต้น

3. Link Analysis (Associations) เป็นการหาความสัมพันธ์ของข้อมูลภายในกลุ่มข้อมูล เพื่อใช้ลักษณะของข้อมูลหนึ่งๆ ในการบอกลักษณะที่จะเกิดขึ้นกับข้อมูลอีกตัวหนึ่งซึ่งอาจจะเป็นการหาความสัมพันธ์ของข้อมูลในกลุ่มเดียวกัน เช่น การระบุว่าในกลุ่มของลูกค้าที่ซื้อนม นั้น จะมีลูกค้า 64% ที่ซื้อขนมปังด้วย หรืออาจจะเป็นการหาความสัมพันธ์ของตัวแปรระหว่างกลุ่มข้อมูลก็ได้ เช่น ในทุกๆ ครั้งที่ดัชนีของตลาดหุ้นหนึ่งลดลง 5% ดัชนีของตลาดหุ้นอื่นจะเพิ่มขึ้น 13% ภายในช่วง 2-6 เดือนหลังจากนั้น เป็นต้น

ลักษณะของการหาความสัมพันธ์นั้นอาจแบ่งได้เป็น 3 กลุ่ม คือ

- การหาความสัมพันธ์ระหว่างข้อมูล (Association discovery)
- การหาความสัมพันธ์ในลักษณะที่เป็นลำดับของข้อมูล (Sequential Pattern discovery)
- การหาความสัมพันธ์ของข้อมูลกับช่วงเวลาใด ๆ (Similar Time Sequence discovery)

4. Deviation Detection เป็นเทคนิคที่ใช้ในการแสดงลักษณะของข้อมูลที่ผิดปกติ หรือผิดไปจากที่คาดไว้ โดยใช้เทคนิคทางด้านสถิติในการหาค่าที่แตกต่างไปจากค่ามาตรฐานแล้วมีการแสดงผล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ในลักษณะที่สามารถทำความเข้าใจและแปลความหมายได้ง่ายขึ้น เช่น การใช้กราฟดูการกระจายของค่าบนกราฟนั้น เป็นต้น

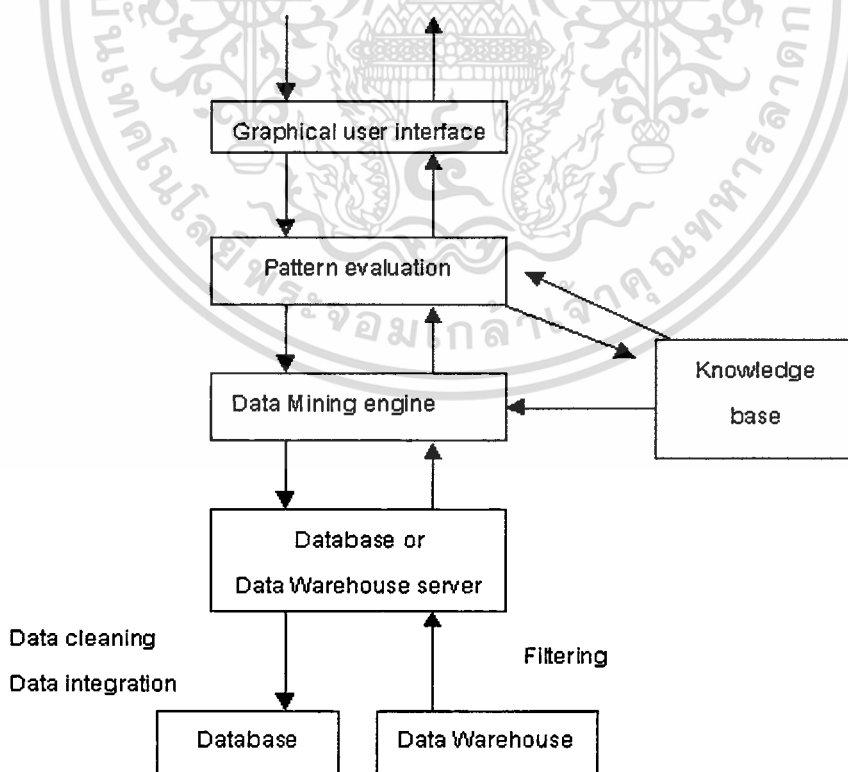
การประยุกต์ใช้ในทางธุรกิจเพื่อพยายามหาสิ่งที่ผิดปกติออกจากกลุ่มของข้อมูลโดยการตรวจสอบค่าเบี่ยงเบน ซึ่งสามารถรองรับการตรวจจับกลโกง เช่น การตรวจสอบลายเซ็นปลอม บัตรเครดิตปลอม การอ้างสิทธิเอาประกัน และการโกงบัตรโทรศัพท์ หรือในเรื่องการควบคุมคุณภาพและการตรวจจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

3.2.5 Interpretation

หลังจากที่การสร้างโมเดลแล้วจำเป็นต้องตรวจสอบผลลัพธ์และตีความหมาย ความถูกต้องที่ตรวจออกมาได้นั้น เป็นชุดตัวอย่างที่ส่งเข้าไป ดังนั้นผลลัพธ์ที่ได้ อาจมีความแปรปรวนหากมีการนำไปใช้กับข้อมูลอื่นๆ

3.2.6 Presentation

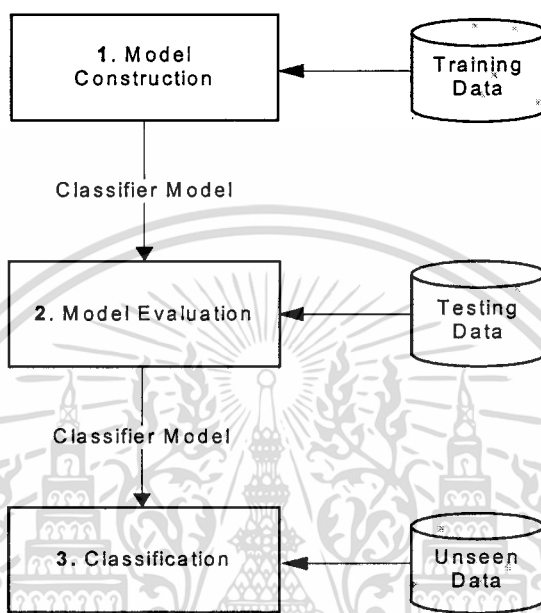
เป็นการแสดงผลการวิเคราะห์โดยอาศัยเครื่องมือที่มีความสามารถและเข้าใจง่าย การแสดงผลอาจจะอยู่ในรูปแบบของ รายงาน ตาราง กราฟ แผนที่หลายมิติ เป็นต้น



เอกสารนี้เป็นเอกสารที่สงวนลิขสิทธิ์ 3.1 แสดงกระบวนการทำดาต้าไมนิ่งโดยทั่วไปนำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.3 การจัดหมวดหมู่ของข้อมูล (Data Classification)

เทคนิคของค้ำไมนึ่งที่สำคัญเทคนิคหนึ่งคือ Data Classification เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่นักวิเคราะห์ได้กำหนดไว้ก่อนแล้ว ตัวอย่างเช่น จัดประเภทของลูกค้ำว่าเชื่อถือได้หรือไม่โดยพิจารณาจากข้อมูลที่มีอยู่



รูปที่ 3.2 กระบวนการ Classification

กระบวนการ Classification นี้แบ่งออกเป็น 3 ขั้นตอนดังนี้

1. Model Construction (Learning) เป็นขั้นการสร้างโมเดลโดยอาศัยการเรียนรู้จากข้อมูลที่ได้กำหนดคลาสไว้เรียบร้อยแล้ว (Training data) ซึ่งโมเดลที่ได้อาจแสดงในรูปของ Decision Tree ที่แต่ละ Node แสดง Attribute, แต่ละกิ่งแสดงผลในการทดสอบ และ Leaf Node แสดงคลาสที่กำหนดไว้ ซึ่ง Decision Tree นี้ง่ายต่อการปรับเปลี่ยนเป็น Classification rule จากตัวอย่างการแบ่งประเภทลูกค้ำ Training Data คือ ข้อมูลของลูกค้ำเก่าที่ทราบทั้งประวัติส่วนตัว รายละเอียดการใช้งาน การชำระค่าบริการ ระยะเวลาที่เป็นลูกค้ำ และข้อมูลความน่าเชื่อถือของลูกค้ำทั้งหมด

2. Model Evaluation (Accuracy) เป็นขั้นการประเมินความถูกต้องโดยอาศัยข้อมูลที่ใช้ทดสอบ (Testing Data) ซึ่งคลาสที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับคลาสที่หาได้จากโมเดลเพื่อทดสอบความถูกต้อง จากตัวอย่าง Testing Data คือข้อมูลลูกค้ำเก่าที่ทราบข้อมูลทั้งหมดเหมือนกับ Training Data แต่ได้แบ่งข้อมูลลูกค้ำมาจำนวนหนึ่งเป็น Testing Data เพื่อใช้ทดสอบความถูกต้องหลังจากสร้างโมเดลด้วย Training Data เรียบร้อยแล้ว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. Model Usage (Classification) เป็นโมเดลสำหรับใช้ กับข้อมูลที่ไม่เคยเห็นมาก่อน (unseen data) โดยจะกำหนดคลาสให้กับออบเจกต์ใหม่ที่ได้มา หรือพยากรณ์ค่าออกมาตามที่ต้องการ จากตัวอย่างเมื่อมีข้อมูลลูกค้าใหม่ (unseen data) ได้แก่ ข้อมูลประวัติส่วนตัว ระยะเวลาที่เป็นลูกค้า และรายละเอียดการใช้งาน และต้องการทราบว่าลูกค้าใหม่ที่เราขายนี้มีความน่าเชื่อถือหรือไม่ จะสามารถหาข้อมูลความน่าเชื่อถือได้โดยนำข้อมูลลูกค้าคนใหม่ไปผ่านโมเดลที่ได้สร้างไว้ เพื่อเปรียบเทียบข้อมูลประวัติต่าง ๆ ของลูกค้าคนใหม่กับ โมเดลของลูกค้าเก่าทั้งหมดที่ แล้วพยากรณ์ออกมาว่าลูกค้าคนใหม่จะอยู่ในกลุ่มที่น่าเชื่อถือหรือไม่จากโมเดลของลูกค้าที่ได้สร้างไว้แล้ว

3.4 การประเมินค่าความแม่นยำของค่าพยากรณ์ (Predictive value) ที่ได้จากแบบจำลองพยากรณ์ (Predictive Model)

การตรวจสอบความแม่นยำของค่าพยากรณ์ที่ได้จากแบบจำลองพยากรณ์นั้นเป็นสิ่งสำคัญเพื่อใช้ในการตรวจสอบและกรองค่าพยากรณ์ที่ผิดแต่ว่าแบบจำลองพยากรณ์ได้ทำนายเป็นไปในทางบวกออกจากค่าพยากรณ์ที่ผิดและค่าพยากรณ์ที่แบบจำลองพยากรณ์ทำนายไปในทางลบที่ถูกต้อง ซึ่งค่าพยากรณ์ที่เราได้จากแบบจำลองพยากรณ์ได้จะสามารถแบ่งออกได้เป็น 4 กรณีดังตารางที่ 3.3

ตารางที่ 3.1 ผลของการประเมินแบบจำลองพยากรณ์

Predict	Positive	Negative
Actual Positive	TRUE positive	FALSE negative
Actual Negative	FALSE positive	TRUE negative

เราสามารถอธิบายค่าพยากรณ์นี้เปรียบเทียบกับ การตรวจสอบวินิจฉัยในการตรวจโรคของแพทย์เพื่อความเข้าใจที่ง่ายขึ้นดังนี้

- 1. True Positive คือ ผลของการพยากรณ์มีค่าเป็น TRUE และผลลัพธ์จริงมีค่าเป็น TRUE นั่นคือแพทย์วินิจฉัยว่าผู้ป่วยมีโรคและผลตรวจสอบก็พบว่าผู้ป่วยมีโรคจริง
- 2. False Positive คือ ผลของการพยากรณ์มีค่าเป็น TRUE และผลลัพธ์จริงมีค่าเป็น FALSE นั่นคือแพทย์วินิจฉัยว่าผู้ป่วยไม่มีโรค แต่ผลการตรวจสอบพบว่าผู้ป่วยมีโรค

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. False Negative คือ ผลของการพยากรณ์มีค่าเป็น FALSE และผลลัพธ์จริงมีค่าเป็น TRUE นั่นคือ แพทย์วินิจฉัยว่าผู้ป่วยมีโรค แต่ผลการตรวจสอบบอกว่าผู้ป่วยไม่มีโรค
4. True Negative คือ ผลของการพยากรณ์มีค่าเป็น FALSE และผลลัพธ์จริงมีค่าเป็น FALSE นั่นคือ แพทย์วินิจฉัยว่าผู้ป่วยไม่มีโรคและผลตรวจสอบพบว่าผู้ป่วยไม่มีโรคเช่นกัน

จากค่าพยากรณ์ทั้ง 4 กรณีดังกล่าวจะเห็นได้ว่าถ้าหากผลจากการประเมินค่าพยากรณ์ที่ได้จากแบบจำลองพยากรณ์เป็นกรณีที่ 2 คือ False Positive นั้นแสดงว่าความแม่นยำของแบบจำลองพยากรณ์ดังกล่าวพยากรณ์ผิดพลาดและมีความแม่นยำน้อยที่สุดเป็นต้น



บทที่ 4

ต้นไม้ช่วยตัดสินใจ

ต้นไม้ช่วยตัดสินใจ (Decision Tree) เป็นอัลกอริทึมที่ใช้เทคนิคการจัดหมวดหมู่และหลักการของการสร้างต้นไม้ช่วยตัดสินใจ (Decision-Tree induction algorithm) ซึ่งเป็นแบบจำลองต้นไม้ที่มีโครงสร้างกฎต่างๆ ขึ้นเพื่อใช้ในการตัดสินใจ โดยปกติมักประกอบด้วยกฎในรูปแบบ “ถ้า เงื่อนไข แล้ว ผลลัพธ์” เช่น

“If Income = High and Married = No THEN Risk = Poor”

“If Income = High and Married = Yes THEN Risk = Good”

อัลกอริทึมนี้ได้รับความนิยม เนื่องจากความไม่ซับซ้อนของอัลกอริทึมทำให้เครื่องมือที่ใช้ในการทำดาต้าไมนิ่งที่วางขายกันอยู่ในท้องตลาดต่างก็ใช้วิธีนี้ข้อดีของวิธีนี้ คือ สามารถตีความและเข้าใจลักษณะของรูปแบบข้อมูล (Pattern) ได้ง่ายเพราะมีการแยกออกเป็นกฎ หรือข้อกำหนดต่างๆ และแบบจำลองการพยากรณ์ในรูปแบบ Decision Tree มีการทำงานแบบ Supervised Learning คือ สร้างแบบจำลองต้นไม้ที่ได้กำหนดหมวดหมู่จากกลุ่มตัวอย่างของข้อมูล แล้วนำไปใช้พยากรณ์กลุ่มของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่

4.1 การสร้างต้นไม้ช่วยตัดสินใจ

อัลกอริทึมพื้นฐานของการสร้างต้นไม้ช่วยตัดสินใจ คืออัลกอริทึม greedy โดยจะสร้างต้นไม้จากบนลงล่างแบบวนซ้ำ (recursive) ด้วยวิธีการแบ่งปัญหาใหญ่ออกเป็นปัญหาย่อย (divide-and-conquer) ดังรูป

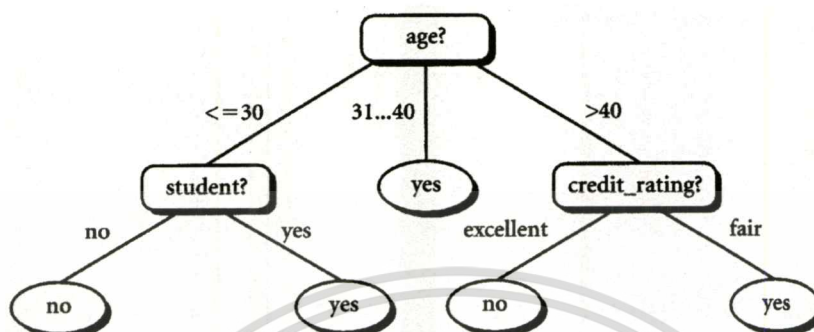
```
(1) create a node  $N$  ;
(2) if samples are all of the same class  $C$  then
(3)   return  $N$  as a leaf node labeled with the class  $C$  ;
(4) if attribute-list is empty then
(5)   return  $N$  as a leaf node labeled with the most
      common class in samples; // majority voting
(6) select test-attribute, the attribute among attribute-
      list with the highest information gain;
(7) label node  $N$  with test-attribute;
(8) for each known value  $a_i$  of test-attribute
(9)   grow a branch from node  $N$  for the condition
      test-attribute =  $a_i$  ;
(10) let  $s_i$  be the set of samples in samples for which
      test-attribute =  $a_i$  ;
(11) if  $s_i$  is empty then
(12)   attach a leaf labeled with the most common class
      in samples ;
(13) else attach the node returned by Generate
      decision tree( $s_i$ , attribute-list, test-attribute);
```

รูปที่ 4.1 อัลกอริทึมพื้นฐานในการสร้างต้นไม้ช่วยตัดสินใจ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า

ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูปแบบของ Decision Tree จะประกอบด้วย Node แรกสุดที่เรียกว่า Root Node จาก Root Node ก็จะแตกออกเป็น Node ลูก และที่ Node ลูกก็จะมีลูกของตัวเองซึ่ง Node ที่ระดับสุดท้ายจะเรียกว่า Leaf Node



รูปที่ 4.2 แสดงตัวอย่างของ Decision Tree

จะเห็นว่าจาก Root Node จนถึง Leaf Node จะมีเพียงเส้นทางเดียวเท่านั้น ซึ่งเส้นทางนี้จะอธิบายถึงกฎที่ใช้สำหรับการจัดหมวดหมู่ของแต่ละกลุ่ม ซึ่งในแต่ละ Leaf Node นั้นอาจเป็นกลุ่มเดียวกัน ซึ่งเกิดจากเหตุผลที่แตกต่างกันได้

วิธีการที่ใช้สร้าง Decision Tree การนำข้อมูลมาสร้าง Tree มีขั้นตอนดังนี้

1. เลือก Attribute ที่สำคัญที่สุดมาแบ่งข้อมูลโดย Attribute นี้จะถูกนำมาสร้างเป็น Root Node โดยจะมี Target Attribute เป็นผลลัพธ์ซึ่งเป็น Leaf Node ถูกกำหนดไว้ก่อน
2. นำค่าที่เป็นไปได้ใน Attribute ที่ถูกเลือกมาแตกออกเป็นกลุ่มของตัวเองที่แตกออกมาจาก Root Node
3. ถ้า Leaf Node มีข้อมูลของหลายคลาสปะปนกันอยู่ ต้องสร้าง Sub Tree เพื่อจำแนกข้อมูลต่อไป โดยเลือก Attribute ที่สำคัญมาทำหน้าที่เป็น Root Node ของ Sub Tree นั้น

4.2 ข้อจำกัดของ Decision Tree

- การแบ่งกลุ่มแบบ Decision Tree กรณีเป็นข้อมูลที่มีค่าต่อเนื่อง เช่น ข้อมูลรายได้ ข้อมูลราคา ต้องทำการแปลงให้อยู่ในช่วงหรือตัดเป็นกลุ่มก่อน

- เมื่อ Algorithm เลือกว่าจะใช้ค่าไหนเป็นตัวแบ่งกลุ่มแล้วก็จะไม่สนใจค่าอื่นที่อาจมีความสำคัญเช่นเดียวกัน

- การจัดการกับข้อมูลที่ไม่ทราบค่า อาจมีผลกระทบกับผลลัพธ์ของ Decision Tree

- Tree ที่มีระดับชั้นมากเกินไปจะทำให้ข้อมูลที่ผ่าน Node แตกออกเป็นชั้นเล็กๆ ซึ่งข้อมูลเหล่านั้นจะไม่มีประโยชน์ในการนำมาใช้ทำการวิเคราะห์

ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ตัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

• ปัญหาเรื่อง Overfitting / Overtraining เกิดจากการที่แบบจำลองได้เรียนรู้เข้าไปถึงรายละเอียดของข้อมูลมากเกินไปจนทำให้เกิด Node ที่เป็นส่วนเฉพาะเจาะจงกับกลุ่มข้อมูลที่ใช้ในการเรียนรู้ ซึ่งจะต้องหาวิธีการในการตัดกิ่งนี้ออกไป หรือกล่าวอีกนัยหนึ่งว่าปัญหาการ Overfitting คือการที่ Model ที่ได้จากการใช้ Training Data set มีค่าความถูกต้องในการบ่งบอกคลาสเป้าหมายสูง แต่เมื่อนำไปใช้กับข้อมูล Test Data set ได้ค่าความถูกต้องต่ำ นั่นคือ ตัวแบบที่ได้เป็นการเรียนรู้ข้อมูลจาก Training Data Set ดีมาก แต่ไม่สามารถนำไปใช้กับข้อมูลที่ไม่เคยพบมาก่อนได้ดี ลักษณะปัญหา Overfitting มักเป็นต้นไม้มันที่มีการแตกกิ่งมากมาย แต่ละกิ่งมีจำนวนข้อมูลอยู่ปริมาณน้อย วิธีการหลีกเลี่ยงปัญหา Overfitting ในต้นไม้มันทำได้ 2 ลักษณะ

1. ตัดแต่งกิ่งระหว่างสร้าง (Pre-Pruning) หยุดการสร้างเมื่อตัววัดการแตกกิ่งมีค่าต่ำกว่าเกณฑ์ที่กำหนด เช่น ข้อมูลในจุดยอดมีน้อยเกินไป

2. ตัดแต่งกิ่งหลังสร้าง (Post-Pruning) การลดจำนวนกิ่ง กำจัดกิ่งโดยยังคงความถูกต้องของการ Classification ในระดับที่ยอมรับได้

4.3 อัลกอริทึม Supervised Learning in Quest (SLIQ)

ปัญหาที่สำคัญเรื่องหนึ่งของการทำดาต้าไมนิ่งโดยใช้อัลกอริทึมที่เกี่ยวกับการจัดหมวดหมู่ คือพื้นที่ของหน่วยความจำ ซึ่งอัลกอริทึมส่วนใหญ่จะถูกออกแบบให้ใช้พื้นที่หน่วยความจำขนาดใหญ่ในการทำงาน ซึ่งทำให้จำนวนชุดของข้อมูลที่ใช้ในการวิเคราะห์จะถูกจำกัดโดยขนาดของหน่วยความจำ ในโครงการพัฒนาระบบงานนี้จะศึกษาเกี่ยวกับอัลกอริทึม “Supervised Learning in Quest” (SLIQ) ซึ่งเป็นอัลกอริทึมที่ใช้ในการจัดหมวดหมู่ของข้อมูลโดยใช้เทคนิคของ Decision Tree

อัลกอริทึม SLIQ เหมาะกับการจัดการกับข้อมูลขนาดใหญ่ ซึ่งสามารถจัดการกับข้อมูลได้ทั้งแบบตัวเลข (Numeric Attributes) และข้อมูลที่จัดเป็นหมวดหมู่ (Categorical Attributes) โดยใช้เทคนิค Pre-sorting data คือการเรียงลำดับข้อมูลพร้อมกันกับการสร้างแบบจำลอง (Tree-growth phase) ที่เรียกว่า Breadth-first growing เพื่อที่จะย้ายข้อมูลบางส่วนลงไปเก็บในดิสก์แทน และในขั้นตอนการปรับแต่งแบบจำลองต้นไม้มัน (Tree-pruning) ก็นำเทคนิค Minimum Description Length Principle (MDL) มาใช้ซึ่งจะมีผลทำให้เพิ่มความถูกต้องของแบบจำลองมีมากยิ่งขึ้น ทำให้สามารถจัดแบ่งกลุ่มข้อมูลได้โดยไม่ต้องคำนึงถึงขนาด หรือชนิดของข้อมูล ซึ่งทำให้เราสามารถนำข้อมูลที่มีเพื่อนำมา สร้างแบบจำลองเพื่อใช้พยากรณ์ได้ดียิ่งขึ้น

4.4 หลักการสร้างต้นไม้ตัดสินใจ (Decision Tree Building)

การสร้างต้นไม้ตัดสินใจนั้นจะแบ่งเป็น 2 ขั้นตอนคือ

1. การสร้างแบบจำลองต้นไม้ (Tree Building) คือ การเลือก Attribute ที่เหมาะสมเพื่อใช้ในการวิเคราะห์ข้อมูลต่อไป และการแบ่งกลุ่มข้อมูลที่ใช้ในการสร้างแบบจำลองต้นไม้ด้วยค่าที่ดีที่สุด (Training Data) อาจจะแบ่งออกเป็น 2 ส่วน หรือมากกว่า 2 ส่วนก็ได้ตาม Attribute ที่กำหนดไว้ การทำงานในขั้นตอนนี้จะเป็นทำวนซ้ำเป็นรอบๆ จนกระทั่งแต่ละข้อมูลตัวอย่างนั้นอยู่ในคลาสเดียวกัน

2. การปรับแต่งแบบจำลองต้นไม้ (Tree Pruning) คือ การสร้างแบบจำลองต้นไม้มักจะพบกับปัญหาข้อมูลผิดพลาด ซึ่งอาจจะเกิดจากการบันทึกข้อมูลผิด ข้อมูลที่ผิดพลาดแบบนี้เรียกว่า ข้อมูลรบกวน (Noise Data) ซึ่งทำให้แบบจำลองต้นไม้ที่สร้างขึ้นจะวิเคราะห์ข้อมูลผิดพลาด และอาจจะสร้างแบบจำลองต้นไม้ที่มีขนาดใหญ่และซับซ้อนเกินไป เนื่องมาจากการพยายามที่จะขยายโครงสร้างเพื่อให้สามารถอธิบายข้อมูลรบกวนเหล่านั้นให้ได้ การวิเคราะห์ข้อมูลที่ดีจึงจะต้องมีเทคนิคที่ตรวจจับและจัดการกับข้อมูลรบกวนได้อย่างมีประสิทธิภาพซึ่งในขั้นตอนนี้จะต้องทำการตรวจสอบแบบจำลองที่ได้สร้างไป และเลือกแบบจำลองต้นไม้ย่อย (Sub Tree) ที่มีข้อผิดพลาดน้อยที่สุด ซึ่งมีวิธีหาค่าผิดพลาดอยู่ 2 วิธีคือ

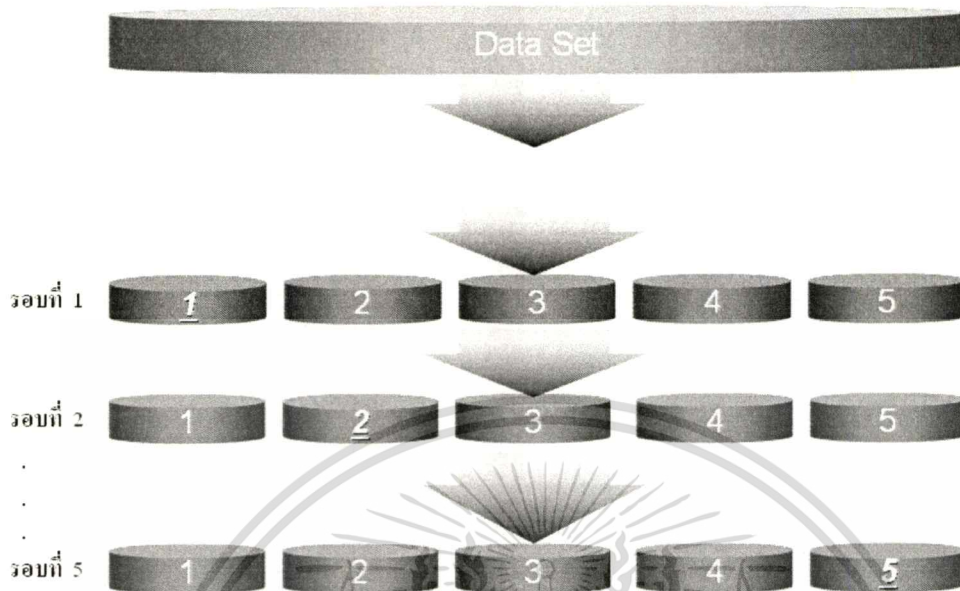
- ใช้ชุดของ Training Data ชุดเดิมวิธีนี้เรียกว่า Cross-Validation โดยนำ Training Data มาแบ่งออกเป็นหลายๆ ตัวอย่างแล้วสร้างแบบจำลองต้นไม้จากตัวอย่างเหล่านี้ จากนั้นก็ใช้แบบจำลองนี้ในการประเมินหาข้อผิดพลาด (Error) กับแบบจำลองที่เราได้สร้างขึ้น ถึงแม้ว่าวิธีการนี้จะทำให้ได้แบบจำลองที่ขนาดเล็กและมีความถูกต้องสูง แต่ว่าต้นทุนในการสร้างแบบจำลองแต่ละแบบจำลองนั้นค่อนข้างสูง ซึ่งทำให้วิธีนี้ไม่ค่อยเหมาะกับชุดของ Training Data ขนาดใหญ่

- ใช้ข้อมูลที่เป็น Training Data ชุดใหม่ วิธีการนี้จะแบ่ง Training Data ออกเป็น 2 ส่วน ส่วนหนึ่งจะใช้ในการสร้างแบบจำลองต้นไม้ อีกส่วนหนึ่งจะใช้ในการปรับแต่ง (Pruning) ซึ่งข้อมูลที่ใช้ในการปรับแต่งก็ควรจะเลือกข้อมูลที่มีการกระจายตามความเป็นจริง เพราะถ้าเลือกข้อมูลขนาดไม่เหมาะสมหรือเลือกข้อมูลไม่ถูก จะมีผลต่อการลดขนาดและความถูกต้องของ Training Data ในการสร้างแบบจำลองต้นไม้ได้

- ตัวอย่างการทำ Cross Validation ด้วยวิธี K-Fold

ในกรณีการทำ K - Fold cross-validation เราจะแบ่งข้อมูลออกเป็น K ชุดเท่าๆกัน และทำการคำนวณค่าความผิดพลาด K รอบ โดยแต่ละรอบการคำนวณข้อมูลชุดหนึ่งจากข้อมูล K ชุดจะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก K - 1 ชุดจะถูกใช้เป็น Training Data

K-fold Cross-Validation



รูปที่ 4.3 แสดงวิธีทำ Cross Validation ด้วยวิธี K-Fold

4.5 การสร้างแบบจำลองต้นไม้ (Tree Building Model) ของอัลกอริทึม SLIQ

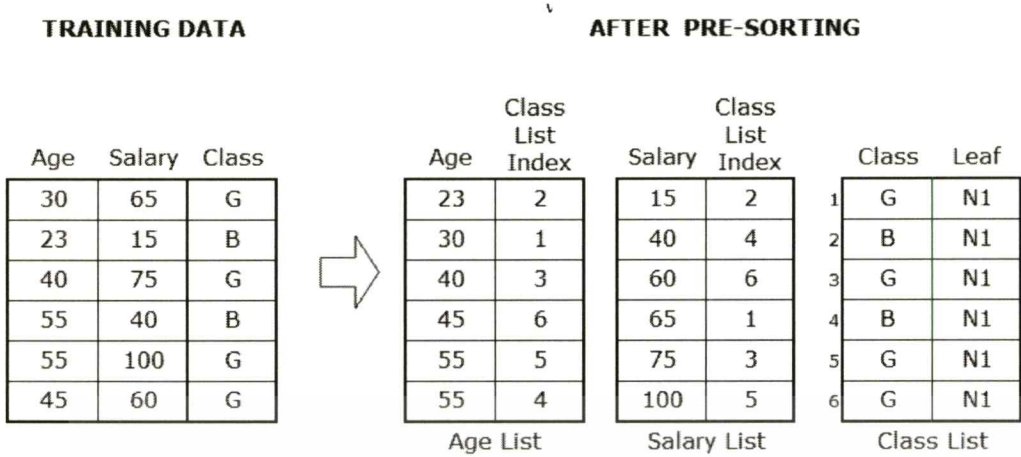
การสร้างแบบจำลองต้นไม้ในอัลกอริทึม SLIQ แบ่งออกเป็น 2 ขั้นตอนดังนี้

1. Pre-Sorting และ Breadth-First Growth เป็นขั้นตอนที่นำข้อมูลมาเรียงลำดับ ซึ่งข้อมูลที่เป็นตัวเลขนั้นการเรียงลำดับข้อมูลถือเป็นสิ่งที่จำเป็นที่สุด ดังนั้นเทคนิคแรกของการทำ Decision Tree โดยใช้อัลกอริทึม SLIQ ก็ต้องนำข้อมูลที่เป็น Numeric มาเรียงลำดับ (การเรียงลำดับนั้นให้เรียงเพียงครั้งแรกครั้งเดียว)

ในการทำงานของ Pre-Sorting นั้นเราต้องทำการจัดโครงสร้างข้อมูลให้เป็นดังนี้

- ทำการกำหนดคลาสสิคส์ (Class List) ขึ้นมาก่อนซึ่งใน Class List นั้นจะประกอบไปด้วย 2 필ด์ดังนี้ คือ Class Label และ Leaf Node ซึ่ง Class Label คือค่าของ Attribute ที่ใช้ในการพยากรณ์ เช่น G = Good, B = Bad

- สร้าง Attribute List ซึ่งใน Attribute List นั้นจะประกอบไปด้วยค่า Attribute และค่าดัชนี (Index) และลำดับที่ I ใน Class List จะมีผลต่อลำดับที่ I ของข้อมูล ในแต่ละ Leaf Node ของ Decision Tree จะแสดงผลของการแบ่งข้อมูลใน Training Set (T) ในการสร้างแบบจำลองต้นไม้ ครั้งแรกจะกำหนดให้ Leaf Node ในคลาสสิคส์ชี้ไปที่ Root Node



รูปที่ 4.4 การสร้างคลาสลิสต์ด้วยวิธีเรียงลำดับก่อน

2. ขั้นตอนการแบ่งโหนด (Node Splits) ในการแบ่งโหนดนั้นจะนำ Gini (Population Diversity) เทคนิคเข้ามาช่วย ซึ่งในการคำนวณสำหรับแต่ละ Attribute นั้นต้องนำการกระจายความถี่ของค่าในแต่ละคลาสมาคิดด้วย ซึ่งก็คือ การทำฮิสโทแกรม (Histogram) ถ้า Attribute เป็นค่า Numeric ตัวฮิสโทแกรมจะมีรูปแบบดังนี้ <class, frequency> ถ้า Attribute เป็น Categorical ตัวฮิสโทแกรมจะมีรูปแบบดังนี้ <Attribute value, class, frequency> เมื่อทำการหา Best Split ได้แล้วก็ให้ทำการสร้าง Tree และทำการปรับค่า Class List

ขั้นตอนการแบ่งโหนดในการสร้างแบบจำลองต้นไม้มีดังนี้

1. เลือก Attribute ที่เหมาะสมเพื่อใช้ในการวิเคราะห์ข้อมูล และทำการแบ่งกลุ่ม Attribute ออกเพื่อหาจุดที่ดีที่สุดเรียกจุดนี้ว่า Best Split ในการเลือกและแบ่ง Attribute ซึ่งเราจะใช้สูตรคำนวณในการหา คือ gini ซึ่งมีสูตรที่ 4.1 ดังนี้

$$gini(T) = 1 - \sum_{j=1}^n p_j^2$$

(4.1)

T: ชุดของข้อมูลที่เก็บตัวอย่างจาก n คลาส

p_j : ความถี่สัมพันธ์ของคลาส j ใน T

ถ้าชุดของข้อมูล T ได้แบ่งเป็น 2 สับเซต ด้วยขนาด N_1 และ N_2 คลาส ค่า $Gini_{split}$ จากจำนวน N คลาสจะเป็นสูตรที่ 4.2 ดังนี้

$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

(4.2)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. สร้างจุดแบ่งโดยใช้ค่า Gini split ที่ต่ำที่สุดมาเป็นตัวกำหนดจุดแบ่ง ซึ่งวิธีการกำหนดจุดแบ่งนั้นจะต้องแบ่งตามประเภทของ Attribute ด้วยดังนี้

- การแบ่งข้อมูลสำหรับ Attribute ที่เป็นตัวเลข (Numeric) เป็นการแบ่งข้อมูลแบบเป็น 2 ทาง (Binary Split) จากรูปแบบ $A \leq v$ โดยที่ v เป็นตัวเลขจำนวนจริงของ Attribute A ซึ่งให้ทำการเรียงข้อมูลที่น่ามาใช้ทดสอบซึ่งขึ้นอยู่กับค่าของ Attribute A ที่พิจารณา เช่นทำการเรียงค่า $v_1, v_2, \dots, v_n$ และเมื่อหาค่า Best Split ได้ว่าเป็นเรคคอร์ดที่ v_i ก็ให้นำค่ากลางระหว่าง $v_i - v_{i+1}$ มาเป็น Best Split
- การแบ่งข้อมูลสำหรับ Attribute ที่แบ่งเป็นหมวดหมู่ (Categorical) เราจะใช้อัลกอริทึม Subsetting ในการแยกประเภทของ Attribute ชนิดนี้ โดยให้ $S(A)$ เป็นชุดของค่าที่เป็นไปได้ใน Attribute A ดังนั้นการแบ่งข้อมูลจะอยู่ในรูปของ $A \in S'$ ซึ่ง $S' \subset S$ ดังนั้นจำนวนของสับเซตที่เป็นไปได้ n จำนวนจะเท่ากับ 2^n สำหรับ Attribute ที่มีค่าเป็นไปได้น n ค่า

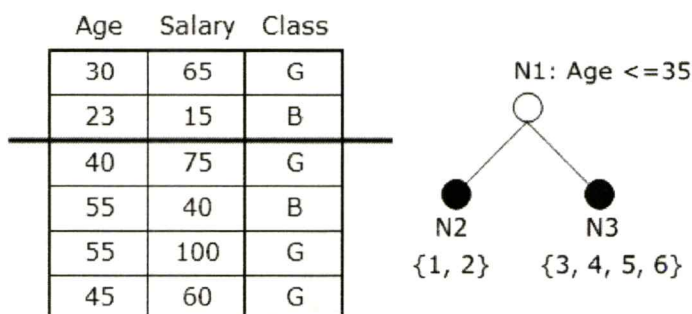
Evaluate Splits()

```

For each attribute list of A DO
  Traverse attributed list of A
  For each value v in the attribute list DO
    Find the corresponding entry in the class
    list, and hence
    the corresponding class and the leaf node
    (say l)
    Update the class histogram in the leaf l
  IF A is a numeric attribute then
    Compute splitting index for test ( $A \leq v$ )
    for leaf l
  IF A is a categorical attribute then
    For each leaf of the tree do
      Find subset of A with best split
  
```

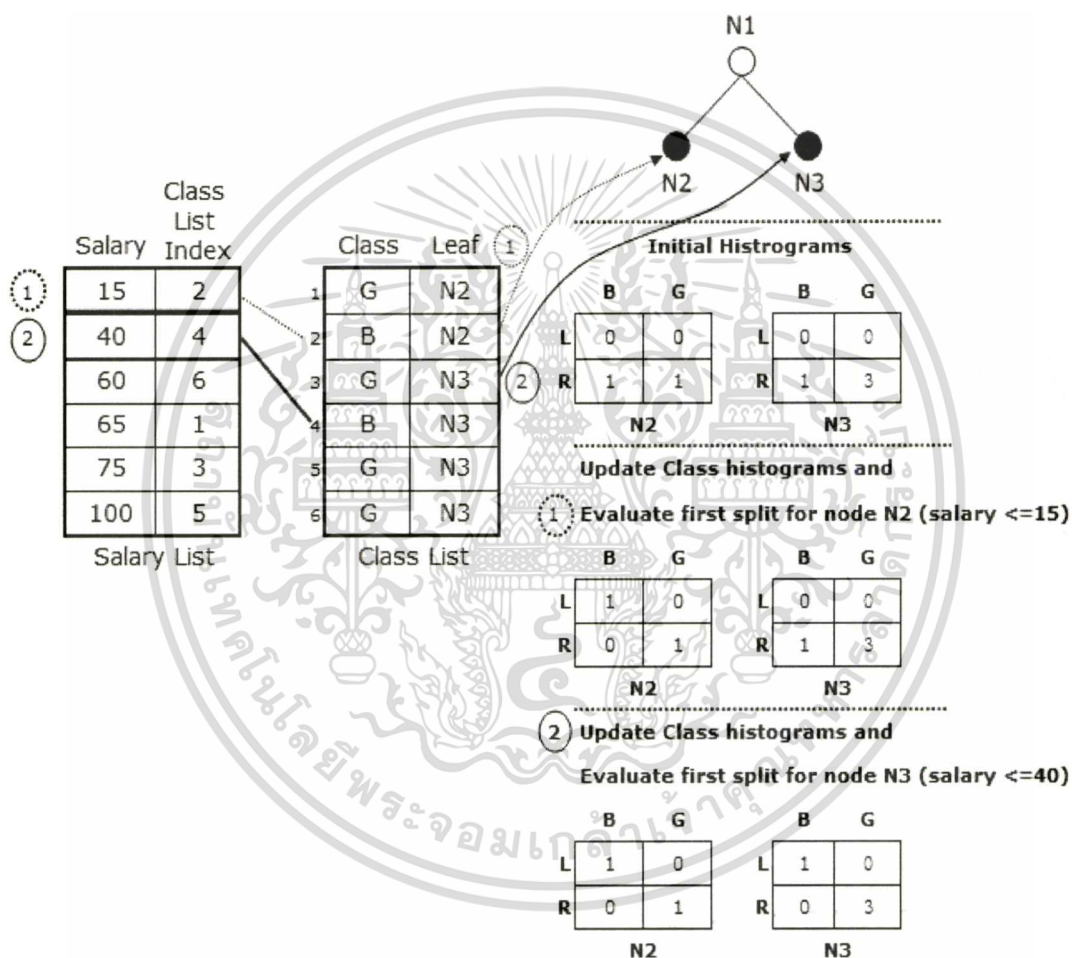
รูปที่ 4.5 อัลกอริทึมการแบ่งโหนดของ SLIQ

จากข้อมูลตัวอย่างรูปที่ 4.4 นั้นสมมติว่าเลือกแอตทริบิวต์ Age เป็น Splitting Index โดยกำหนดค่าเป็น $\text{Age} \leq 35$ จะทำให้สามารถแบ่งโหนดออกได้ตามรูปที่ 4.6



เอกสารนี้เป็นเอกสารที่สงวนลิขสิทธิ์และใช้เฉพาะเพื่อการศึกษานี้เท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
 รูปที่ 4.6 การแบ่งโหนดโดยใช้แอตทริบิวต์ Age เป็น Splitting Index
 ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หลังจากแบ่งโหนดในรอบแรกด้วยแอททริบิวต์ Age แล้วจะเห็นว่าข้อมูลของโหนด N2, N3 นั้นยังสามารถจัดกลุ่มออกให้เป็นกลุ่มได้อีกจึงต้องทำการแบ่งโหนดย่อยโดยใช้แอททริบิวต์ Salary เป็น Splitting Index จากข้อมูลตัวอย่างดังรูป 4.7 โดยเริ่มแรกจะเห็นว่าค่า 15 เป็นค่าแรกใน Salary list ซึ่งอยู่ในโหนด N2 ดังนั้นจึงให้ค่า $\text{Salary} \leq 15$ เป็นค่าในการแบ่งโหนด N2 และเช่นเดียวกันกับค่า 40 ซึ่งเป็นค่าที่อยู่ในโหนด N3 ดังนั้นจึงให้ค่า $\text{Salary} \leq 40$ เป็นค่าในการแบ่งโหนด N3



รูปที่ 4.7 การแบ่งโหนดของต้นไม้ตัดสินใจโดยใช้แอททริบิวต์ Salary เป็น Splitting Index

แต่จากการจากการแบ่งโหนด N2 ด้วยค่า $\text{Salary} \leq 15$ นั้นจะเห็นว่าจะได้ผลลัพธ์ที่มีเหมือนกันกับการใช้ $\text{Salary} \leq 40$ ในการแบ่งโหนด ดังนั้นจึงต้องมีการปรับปรุงค่าที่ใช้ในการแบ่งโหนดใหม่อีกครั้งโดย N2 จะใช้ค่า $\text{Salary} \leq 40$ ในการแบ่งโหนดนั่นเอง โดยจะใช้อัลกอริทึมในการปรับปรุงคลาสสิคส์ตามรูปที่ 4.8

Update Labels()

For each attribute A used in a split **DO**
 traverse attributed list of A

For each value v in the attribute list **DO**
 find the corresponding entry in the class
 list e,

find the new class c to which V belongs by
 applying the splitting test at node
 referenced from e

Update the class label for e to c

Update node referenced in e to the child
 corresponding to the class c

รูปที่ 4.8 อัลกอริทึมการปรับปรุงคลาสสถิติของต้นไม้ตัดสินใจ

หลังจากนั้นจะต้องทำการคำนวณหาจุดแบ่งของโหนด N3 โดยใช้ gini split index ซึ่งแสดง
 ดังรูปที่ 4.9 ซึ่งจากการคำนวณนั้นจะเป็นว่าค่า gini split index ของค่า Salary ≤ 50 นั้นต่ำที่สุด
 ดังนั้นจึงใช้ค่าดังกล่าวเป็นจุดแบ่งสำหรับโหนด N3 แล้วทำการปรับปรุงคลาสสถิติให้เหมาะสมอีก
 ครั้งดังรูปที่ 4.10 จึงได้ต้นไม้ตัดสินใจที่สมบูรณ์สำหรับข้อมูลชุดนี้

Salary Class		≤ 35		≤ 50		≤ 70		≤ 80		≤ 120	
		N2	N3	N2	N3	N2	N3	N2	N3	N2	N3
75	G										
40	B										
100	G										
60	G										



≤ 50

$$gini(T) = 1 - \sum p_j^2$$

$$Gini(N2) = 1 - (0/1)^2 - (1/1)^2 = 0$$

$$Gini(N3) = 1 - (3/3)^2 - (0/1)^2 = 0$$

$$Gini(split) = 1/4 * 0 + 3/4 * 0 = 0$$

≤ 70

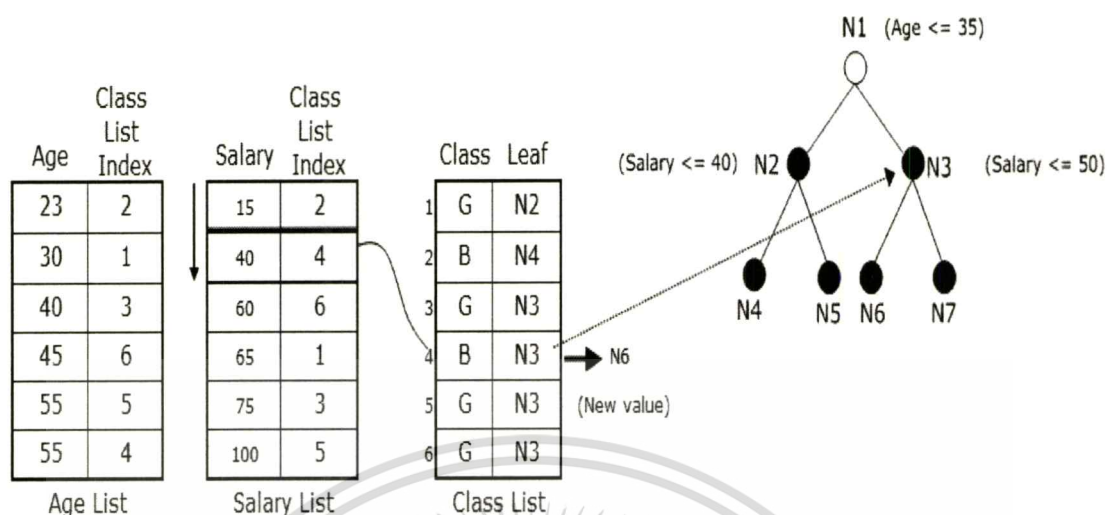
$$Gini(N2) = 1 - (1/2)^2 - (1/2)^2 = 0$$

$$Gini(N3) = 1 - (2/2)^2 - (0/2)^2 = 0$$

$$Gini(split) = 2/4 * 1/2 + 2/4 * 0 = 1/4 = 0.25$$

รูปที่ 4.9 การแบ่งโหนดของต้นไม้ตัดสินใจย่อย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
 ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.10 การปรับปรุงคลาสสิคส์ของต้นไม้ตัดสินใจ

4.6 การขยายความสามารถของต้นไม้ตัดสินใจ

- นำต้นไม้ตัดสินใจไปใช้กับลักษณะประจำที่มีค่าต่อเนื่อง โดยการเปลี่ยนให้เป็นค่าไม่ต่อเนื่องระหว่างการสร้างต้นไม้ หรือการทำการก่อนการสร้างต้นไม้
- จัดการกับลักษณะประจำที่ข้อมูลขาดหายไป
 - ใช้ค่าที่ปรากฏบ่อยที่สุดในลักษณะประจำ
 - คำนวณค่าความน่าจะเป็น แล้วใช้การกระจายดังกล่าวกำหนดค่าที่หายไป
- สร้างลักษณะประจำที่มักไม่ถูกใช้สร้างเป็นลักษณะประจำใหม่
 - รวมค่าในลักษณะประจำที่มักไม่ถูกใช้สร้างเป็นลักษณะประจำใหม่
 - ช่วยลด การแตกแยกย่อย (fragmentation), การเกิดซ้ำ (repetition), และการมีข้อมูลซ้ำซ้อน (replication)

บทที่ 5

การออกแบบและการพัฒนาโปรแกรม

ในการศึกษาโครงการนี้ได้นำทฤษฎีแนวคิดและขั้นตอนของการทำค้ำไม้หนึ่งด้วยเทคนิคจัดหมวดหมู่ของข้อมูลโดยใช้อัลกอริทึม SLIQ และนำข้อมูลจากการทำค้ำไม้หนึ่งที่ได้มานำเสนอในรูปแบบของต้นไม้ช่วยตัดสินใจ แล้วนำมาออกแบบ และพัฒนาเป็นโปรแกรมประยุกต์สำหรับช่วยวิเคราะห์ความเสี่ยงของลูกค้าโทรศัพท์เคลื่อนที่เพื่อช่วยลดความเสี่ยงเครดิตของลูกค้า ทั้งนี้ยังมีผลช่วยให้สามารถลดเวลา ลดต้นทุนและค่าใช้จ่ายในการติดตามหนี้ของลูกค้าอีกด้วย โดยการศึกษาดำเนินการมีขั้นตอนดังนี้

5.1 การศึกษาทฤษฎีที่เกี่ยวข้อง

ในการพัฒนาระบบนี้มีทฤษฎีที่ผู้พัฒนาต้องทราบดังต่อไปนี้

1. การบริหารลูกค้าสัมพันธ์ (Customer Relationship Management) หรือ CRM โดยเฉพาะหัวข้อเรื่องการลดความเสี่ยงเครดิตและการติดตามหนี้ อีกทั้งการประเมินพฤติกรรมด้านสินเชื่อของธุรกิจการเงิน
2. ทฤษฎีของการทำค้ำไม้หนึ่งด้วยเทคนิคจัดหมวดหมู่ของข้อมูล
3. อัลกอริทึม SLIQ และการนำเสนอข้อมูลในรูปแบบของต้นไม้ตัดสินใจ
4. ความรู้ความเข้าใจการพัฒนาโปรแกรมและภาษาโปรแกรมประยุกต์โดยใช้ภาษา Visual Basic 6.0 และ Microsoft Access

5.2 การศึกษาความเป็นไปได้ และวิเคราะห์ออกแบบระบบ

จากการศึกษาความเป็นไปได้และการออกแบบระบบนั้นได้ออกแบบให้ระบบมีความสามารถทำงานหลักดังต่อไปนี้

1. กำหนดให้ระบบสามารถนำข้อมูลเข้า (Input) จากฐานข้อมูล Microsoft Access
2. กำหนดให้ระบบผู้ใช้สามารถเลือกตัวแปรที่จะใช้ในการวิเคราะห์ที่กำหนดไว้เองได้ โดยเลือกได้สูงสุด 10 ตามที่กำหนดเอาไว้
3. กำหนดให้ระบบสามารถกำหนดจำนวนระดับและจำนวนโหนดของแบบจำลองต้นไม้ตัดสินใจที่ต้องการจะสร้างเองได้
4. กำหนดให้ระบบสามารถกำหนดและแสดงค่าความถูกต้องของแบบจำลองต้นไม้ตัดสินใจได้
5. กำหนดให้ระบบสามารถแสดงผลแบบ Tree Model และ Classification rule ได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

- 6. กำหนดให้ระบบสามารถจัดเก็บแบบจำลองต้นไม้ตัดสินใจที่ได้สร้างไว้เพื่อนำมาใช้ในการวิเคราะห์ข้อมูลภายหลังได้
- 7. กำหนดให้ระบบสามารถจัดเก็บผลการวิเคราะห์ของข้อมูลเพื่อนำมาสร้างรายงานเพื่อเรียกดูในภายหลังได้

5.3 การจัดเตรียมและจัดเก็บข้อมูลในการวิเคราะห์

ข้อมูลเพื่อนำมาใช้ในการวิเคราะห์นั้นเป็นข้อมูลของลูกค้าแบบจดทะเบียนรายเดือนในระบบของผู้ให้บริการเครือข่ายโทรศัพท์เคลื่อนที่ซึ่งเป็นกลุ่มของลูกค้าที่มีลักษณะการใช้งานโทรศัพท์แบบใช้ก่อนจ่ายทีหลัง (Post Paid) โดยมีการจัดเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมและปรับปรุงให้ข้อมูลมีความถูกต้องก่อนนำมาใช้งาน การปรับแต่งข้อมูลหรือตัดทิ้งข้อมูลที่ไม่ถูกต้องออกนั้นเป็นการทำความสะอาดข้อมูลเพื่อทำให้ข้อมูล (Training Data) ที่จะนำมาใช้ในการสร้างแบบจำลองพยากรณ์ถูกต้องนั่นเอง ซึ่งจะมีผลทำให้เราได้แบบจำลองที่ดีอีกด้วย

5.3.1 การจัดเตรียมข้อมูลในการวิเคราะห์

การจัดเตรียมข้อมูลนี้จะมีการคัดเลือกแอททริบิวต์ของข้อมูลที่สำคัญที่มีแนวโน้มว่าจะมีความเกี่ยวข้องกับความเสี่ยงของลูกค้าโทรศัพท์เคลื่อนที่ที่เกิดขึ้นเพื่อนำมาใช้ในการวิเคราะห์และประเมินความเสี่ยงจำนวน 11 แอททริบิวต์ โดยแสดงในตารางที่ 5.1

ตารางที่ 5.1 แสดงแอททริบิวต์ที่ใช้ในการวิเคราะห์และประเมินความเสี่ยง

ลำดับที่	ชื่อ	รายละเอียด	ชนิดของข้อมูล
1.	AGE	อายุ	Continuous Value*
2.	EDUCATION	ประวัติการศึกษา [1 = ปริญญาเอก] [2 = ปริญญาโท] [3 = ปริญญาตรี] [4 = อนุปริญญา] [5 = อื่นๆ]	Discrete Value**
3.	OCCUPATION	อาชีพ [1 = ข้าราชการ, รัฐวิสาหกิจ, ลูกจ้างประจำ]	Discrete Value

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับใช้เพื่อการศึกษาเท่านั้น ไม่อนุญาตให้เผยแพร่โดยไม่ได้รับอนุญาต
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 5.1 (ต่อ)

ลำดับที่	ชื่อ	รายละเอียด	ชนิดของข้อมูล
3.	OCCUPATION	[2 = ลูกจ้างบริษัท, ห้างร้าน, สำนักงาน] [3 = ผู้ประกอบอาชีพอิสระอื่นๆ ผู้ประกอบการ, เจ้าของ, กรรมการ, หุ้นส่วน] [4 = ลูกจ้าง, กลุ่มบุคคลในสถาบันที่ไม่ แสวงหากำไร] [5 = อื่นๆ]	Discrete Value
4.	POSITION	ตำแหน่ง [1 = ผู้บริหารระดับสูง-กลาง, ข้าราชการระดับสูง-กลาง] [2 = รองผู้จัดการ, หัวหน้า, ผู้ช่วย] [3 = ข้าราชการทั่วไป, พนักงานทั่วไป] [4 = ผู้ประกอบอาชีพอิสระ, รับจ้าง] [5 = อื่นๆ]	Discrete Value
5.	EXPERIENCE	ประสบการณ์ในการทำงาน	Continuous Value
6.	INCOME	รายได้ต่อเดือน	Continuous Value
7.	MARITAL STATUS	สถานะสมรส [0 = โสด] [1 = สมรส] [2 = อื่นๆ]	Discrete Value
8.	CHILD	จำนวนบุตร	Continuous Value
9.	CREDIT LIMIT	จำนวนเงินจำกัด	Continuous Value
10.	PAYMENT	จำนวนเงินที่ชำระค่าบริการ	Continuous Value
11.	CREDIT STATUS	สถานะเครดิต [P = Pass] [N = Not Pass]	Discrete Value

หมายเหตุ

* Continuous Value หมายถึงข้อมูลที่เป็นตัวเลขต่อเนื่อง เช่น รายได้, ยอดขาย, จำนวนครั้ง เป็นต้น

** Discrete Value หมายถึงข้อมูลที่ไม่ต่อเนื่องซึ่งเป็นข้อมูลที่สามารถจัดกลุ่มหรือประเภทข้อมูลได้
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้คัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.3.2 การจัดเก็บข้อมูลที่ได้จากการวิเคราะห์

การจัดเก็บข้อมูลที่ได้จากการวิเคราะห์ด้วยแบบจำลองนี้เพื่อจะนำมาใช้ในเรียกดูหรือทำรายงานในภายหลัง โดยแสดงแอททริบิวต์ในตารางที่ 5.2

ตารางที่ 5.2 แสดงแอททริบิวต์ที่เก็บผลจากการวิเคราะห์และประเมินความเสี่ยง

ลำดับที่	ชื่อ	คำอธิบาย	ชนิดของข้อมูล
1	No.	ลำดับที่	NUMBER
2	REF_DATE	วันที่	TEXT
3.	REF_TIME	เวลา	TEXT
4.	AGE	อายุ	TEXT
5.	EDUCATION	ประวัติการศึกษา [1 = ปริญญาเอก] [2 = ปริญญาโท] [3 = ปริญญาตรี] [4 = อนุปริญญา] [5 = อื่นๆ]	TEXT
6.	OCCUPATION	อาชีพ [1 = ข้าราชการ, รัฐวิสาหกิจ, ลูกจ้างประจำ] [2 = ลูกจ้างบริษัท, ห้างร้าน, สำนักงาน] [3 = ผู้ประกอบอาชีพอิสระอื่นๆ ผู้ประกอบการ, เจ้าของ, กรรมการ, หุ้นส่วน] [4 = ลูกจ้าง, กลุ่มบุคคลในสถาบันที่ไม่ แสวงหากำไร] [5 = อื่นๆ]	TEXT
7.	POSITION	ตำแหน่ง [1 = ผู้บริหารระดับสูง-กลาง, ข้าราชการระดับสูง-กลาง] [2 = รองผู้จัดการ, หัวหน้า, ผู้ช่วย] [3 = ข้าราชการทั่วไป, พนักงานทั่วไป]	TEXT

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้เผยแพร่หรือใช้ประโยชน์ด้านการค้า

ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 5.2 (ต่อ)

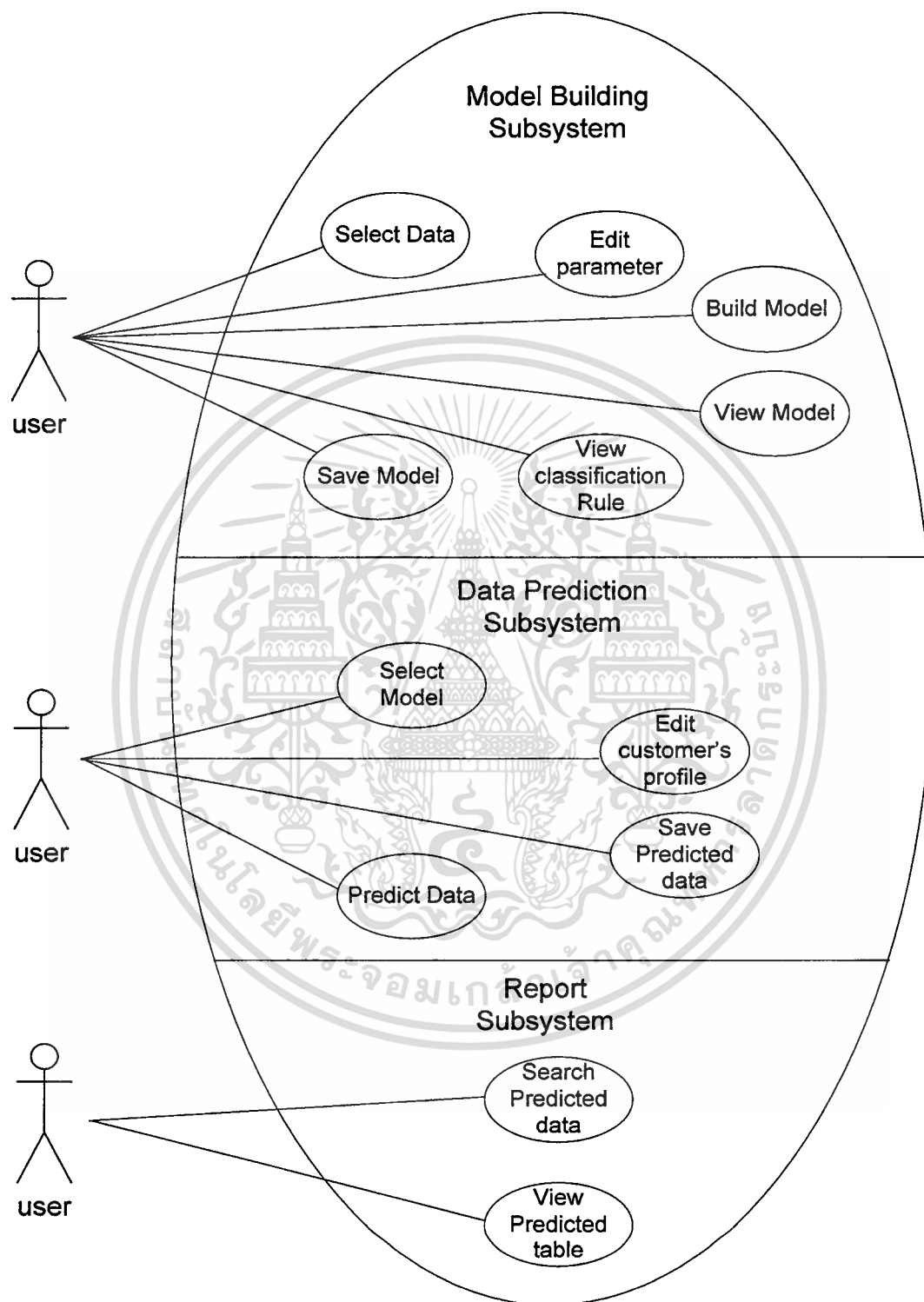
ลำดับที่	ชื่อ	คำอธิบาย	ชนิดของข้อมูล
7.	POSITION	[4 = ผู้ประกอบอาชีพอิสระ, รับจ้าง] [5 = อื่นๆ]	TEXT
8.	EXPERIENCE	ประสบการณ์ในการทำงาน	TEXT
9.	INCOME	รายได้ต่อเดือน	TEXT
10.	MARITAL STATUS	สถานะสมรส [0 = โสด] [1 = สมรส] [2 = อื่นๆ]	TEXT
11.	CHILD	จำนวนบุตร	TEXT
12.	CREDIT LIMIT	จำนวนเงินจำกัด	TEXT
13.	PAYMENT	จำนวนเงินที่ชำระค่าบริการ	TEXT
14.	CREDIT STATUS	สถานะเครดิต [P = Pass] [N = Not Pass]	TEXT
15.	MODEL	แบบจำลอง	TEXT

5.4 การออกแบบระบบ

ในการออกแบบระบบสำหรับโปรแกรมประยุกต์สำหรับช่วยวิเคราะห์ความเสี่ยงของลูกค้า โทรศัพท์เคลื่อนที่เพื่อช่วยลดความเสี่ยงเครดิตของลูกค้า ใช้เครื่องมือในการวิเคราะห์และออกแบบ มีดังต่อไปนี้

- 1. แบบจำลอง Use Case Diagram
- 2. แบบจำลอง Use Case Description
- 3. แบบจำลอง Activity Diagram
- 4. แบบจำลอง Class Diagram

5.4.1 Use Case Diagram ของระบบ



รูปที่ 5.1 Use Case Diagram แสดงความสัมพันธ์ระหว่างผู้ใช้และฟังก์ชันต่างๆ ของระบบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.4.2 Use Case Description ของระบบ

Use Case Description สำหรับอธิบายการทำงานฟังก์ชันย่อยๆ ของระบบโดยคร่าวๆ โดยโปรแกรมประยุกต์ที่ออกแบบจะมีฟังก์ชันการทำงานดังนี้

ตารางที่ 5.3 แสดง Use Case Description ของ Select Data

Use Case Name:	1. Select Data	
Scenario:	เลือกข้อมูลที่ต้องการสร้างแบบจำลองพยากรณ์จากฐานข้อมูล	
Triggering Event:	ผู้ใช้งานกดปุ่มคำสั่ง select data	
Brief Description:	กดปุ่มคำสั่ง select data เพื่อเลือกข้อมูลที่ต้องการใช้ในการสร้างแบบจำลอง	
Actors:	ผู้ใช้งาน	
Related Use Cases:	-	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	-	
Postconditions:	ข้อมูลจากฐานข้อมูลตามแอททริบิวต์ที่ต้องการใช้สร้างแบบจำลอง	
Flow of Events:	Actor	System
	1. ผู้ใช้งานทำการเลือกข้อมูลที่ต้องการ โดยการคลิกปุ่มคำสั่ง SELECT DATA 2. ผู้ใช้งานเลือกค่าแอททริบิวต์ที่ต้องการจากช่อง Available Attribute 3. ผู้ใช้งานเปลี่ยนแปลงค่าแอททริบิวต์ที่ไม่ต้องการออกจากช่อง Selected Attribute 4. กดปุ่มคำสั่ง OK	1.1 ระบบเปิดหน้าต่าง Connect to database ขึ้นมา 2.1 ระบบจะย้ายค่าที่ผู้ใช้เลือกไว้จากช่อง Available Attribute ไปแสดงในช่อง Selected Attribute แทน 3.1 ระบบจะย้ายค่าที่ผู้ใช้เลือกไว้จากช่อง Selected Attribute แสดงไว้ในช่อง Available Attribute เหมือนเดิม 4.1 ระบบจะเก็บค่าข้อมูลในช่อง Selected Data เพื่อนำไปใช้ในการคำนวณต่อไป
Exception Conditions:	-	

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 5.4 แสดง Use Case Description ของ Edit Parameter

Use Case Name:	2. Edit Parameter	
Scenario:	ผู้ใช้งานใส่ค่าเพื่อกำหนดลักษณะของแบบจำลองต้นไม้ที่ต้องการ	
Triggering Event:	-	
Brief Description:	ใส่ค่าตัวเลขที่ต้องการเพื่อกำหนดจำนวนโหนดของแบบจำลองต้นไม้	
Actors:	ผู้ใช้งาน	
Related Use Cases:	1.Select Data, 3.Build Model	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	-	
Postconditions:	แบบจำลองต้นไม้ที่ผู้ใช้กำหนด	
Flow of Events:	Actor	System
	1. ผู้ใช้งานใส่ค่าตัวเลขเพื่อกำหนดจำนวนโหนดและจำนวนระดับของแบบจำลองต้นไม้	1.1 ระบบจะเก็บค่า Parameter ที่ผู้ใช้งานต้องการเพื่อนำไปใช้ในการสร้างแบบจำลองต้นไม้ต่อไป
Exception Conditions:	1. ไม่ควรเว้นเป็นค่าว่าง เนื่องจากเป็นข้อมูลที่จำเป็นต่อการคำนวณของระบบ แต่ถ้าเว้นว่างจะทำให้ระบบใช้เวลาในการประมวลผลนานมาก 2. ไม่ควรใส่เป็นตัวอักษร 3. ไม่ควรใส่ค่าตัวเลขที่มากเพราะมีผลต่อความถูกต้อง	

ตารางที่ 5.5 แสดง Use Case Description ของ Build Model

Use Case Name:	3. Build Model	
Scenario:	สร้างแบบจำลองตามที่ผู้ใช้งานกำหนด	
Triggering Event:	ผู้ใช้งานกดปุ่มคำสั่ง Build Model	
Brief Description:	ผู้ใช้งานสามารถสั่งให้เริ่มการสร้างแบบจำลองโดยการกดปุ่มคำสั่ง Build Model	
Actors:	ผู้ใช้งาน	
Related Use Cases:	1.Select Data, 2.Edit Parameter	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องมีการเลือกข้อมูลและใส่ค่า parameter ต่างๆ ให้ครบถ้วนไว้ก่อน	
Postconditions:	แบบจำลองต้นไม้	
Flow of Events:	Actor	System
	1. ผู้ใช้งานกดปุ่ม Build Model 2. หากผู้ใช้งานไม่พอใจในแบบจำลองต้นไม้ที่สร้างขึ้นมานั้น ผู้ใช้งานสามารถเคลียร์ค่าของข้อมูลและ parameter ต่างๆ ก่อนหน้าเพื่อจะเลือกข้อมูลและใส่ parameter ใหม่ที่ต้องการเพื่อสร้างแบบจำลองใหม่อีกครั้ง	1.1 ระบบจะนำค่า parameter ต่างๆ ที่ผู้ใช้งานได้เลือกและกำหนดเอาไว้ไปใช้ในการประมวลผลออกมาเพื่อสร้างเป็นแบบจำลองต้นไม้ 1.2 ระบบจะเคลียร์ค่าต่างๆ ที่ใช้ในการสร้างแบบจำลองและการประมวลผล และผลลัพธ์ในการประมวลผลก่อนหน้านี้ เช่น ค่าแอททริบิวต์ต่างๆ, แบบจำลองต้นไม้, กฎการตัดสินใจ เป็นต้น
Exception Conditions:		

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ตัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 5.6 แสดง Use Case Description ของ View Model

Use Case Name:	4. View Model	
Scenario:	แสดงแบบจำลองที่สร้างขึ้นในรูปแบบต้นไม้ตัดสินใจ	
Triggering Event:	ผู้ใช้เลือก Decision Tree Tab	
Brief Description:	ผู้ใช้งานสามารถเลือกดูแบบจำลองที่สร้างขึ้นในรูปแบบต้นไม้ตัดสินใจ โดยการเลือก Decision Tree Tab	
Actors:	ผู้ใช้งาน	
Related Use Cases:	3. Build Model	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องมีหน้าจอสำหรับแสดงผล	
Postconditions:	แบบจำลองที่สร้างขึ้นในรูปแบบต้นไม้ตัดสินใจ	
Flow of Events:	Actor	System
	1. ผู้ใช้งานคลิกเลือก Decision Tree Tab	1.1 ระบบจะแสดงหน้าจอย่อยแสดงแบบจำลองต้นไม้ตัดสินใจที่ได้จากการประมวลผล
Exception Conditions:	-	

ตารางที่ 5.7 แสดง Use Case Description ของ View Classification Rule

Use Case Name:	5. View Classification Rule	
Scenario:	แสดงแบบจำลองที่สร้างขึ้นในรูปแบบกฎการตัดสินใจ	
Triggering Event:	ผู้ใช้เลือก Classification Rule Tab	
Brief Description:	ผู้ใช้งานสามารถเลือกดูแบบจำลองที่สร้างขึ้นในแบบกฎการตัดสินใจ โดยการเลือก Classification Rule Tab	
Actors:	ผู้ใช้งาน	
Related Use Cases:	3. Build Model	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องมีหน้าจอสำหรับแสดงผล	
Postconditions:	แบบจำลองที่สร้างขึ้นในรูปแบบกฎการตัดสินใจ	
Flow of Events:	Actor	System
	1. ผู้ใช้งานคลิกเลือก Classification Rule Tab	1.1 ระบบจะแสดงหน้าจอย่อยแสดงแบบจำลองต้นไม้ตัดสินใจที่ได้จากการประมวลผล
Exception Conditions:	-	

ตารางที่ 5.8 แสดง Use Case Description ของ Save Model

Use Case Name:	6. Save Model	
Scenario:	การเก็บบันทึกผลแบบจำลองต้นไม้	
Triggering Event:	ผู้ใช้งานกดปุ่มคำสั่ง Save Model	
Brief Description:	ผู้ใช้งานสามารถเก็บบันทึกผลแบบจำลองต้นไม้แล้วหลังจากการประมวลผลเพื่อนำไปใช้งานได้ในภายหลัง	
Actors:	ผู้ใช้งาน	
Related Use Cases:	3. Build Model	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องมีแบบจำลองที่ผ่านการประมวลผลออกมาแล้ว	
Postconditions:	-	
Flow of Events:	Actor	System
	1. ผู้ใช้งานคลิกคำสั่ง Save Model 2. เลือก directory ที่ต้องการเก็บไว้ในหน่วยความจำสำรอง 3. กดคำสั่ง OK เพื่อยืนยันการ Save Model	1.1 ระบบแสดงหน้าจอข้อย่อย Save file 2.1 ระบบแสดง directory ต่างๆของผู้ใช้ที่ต้องการเก็บข้อมูล 3.1 ระบบทำการเก็บบันทึกข้อมูลโดยไฟล์จะมีนามสกุลเป็น .tree
Exception Conditions:	นามสกุลของไฟล์ที่ต้องการจัดเก็บต้องเป็น .tree เท่านั้น	

ตารางที่ 5.9 แสดง Use Case Description ของ Select Model

Use Case Name:	7. Select Model	
Scenario:	การเลือกแบบจำลองต้นไม้ตัดสินใจจากการบันทึกเพื่อใช้งาน	
Triggering Event:	ผู้ใช้งานกดปุ่มคำสั่ง Select Model	
Brief Description:	ผู้ใช้งานเลือกแบบจำลองต้นไม้ตัดสินใจจากการบันทึกไว้ก่อนหน้านี้เพื่อนำมาใช้ในการประเมินข้อมูลใหม่	
Actors:	ผู้ใช้งาน	
Related Use Cases:	6. Save Model	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องมีแบบจำลองต้นไม้ตัดสินใจที่ได้จัดเก็บไว้ในหน่วยความจำสำรอง	
Postconditions:	แบบจำลองต้นไม้ตัดสินใจที่ผู้ใช้ต้องการ	
Flow of Events:	Actor	System
	1. ผู้ใช้งานเลือกปุ่มคำสั่ง Select Model	1.1 ระบบแสดงหน้าจอย่อย Open file
	2. เลือก directory ที่ต้องการนำมาใช้จากหน่วยความจำสำรอง	2.1 ระบบแสดง directory ต่างๆ ของที่ผู้ใช้ต้องการเลือกข้อมูล
	3. กดปุ่มคำสั่ง Open เพื่อยืนยันการ Select Model	3.1 ระบบทำการแสดงข้อมูล โดยไฟล์จะมีนามสกุลเป็น .tree
		4.1 ระบบทำการแสดงเป็นแบบจำลองต้นไม้ตัดสินใจ
Exception Conditions:	นามสกุลของไฟล์ที่ต้องการจัดเก็บต้องเป็น .tree เท่านั้น	

ตารางที่ 5.10 แสดง Use Case Description ของ Edit Customer’s profile

Use Case Name:	8. Edit Customer’s profile	
Scenario:	ใส่ข้อมูลของลูกค้ารายใหม่ที่ต้องการนำมาวิเคราะห์	
Triggering Event:	-	
Brief Description:	ผู้ใช้งานใส่ข้อมูลของลูกค้ารายใหม่ที่ต้องการจะนำมาวิเคราะห์ด้วยแบบจำลองต้นไม้ตัดสินใจที่ได้เลือกเอาไว้	
Actors:	ผู้ใช้งาน	
Related Use Cases:	7. Select Model	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	แบบจำลองต้นไม้ไม่ได้เลือกเอาไว้ก่อนหน้านี้	
Postconditions:	ผลการวิเคราะห์ข้อมูลด้วยแบบจำลองต้นไม้	
Flow of Events:	Actor	System
	1. ผู้ใช้งานใส่ค่าต่างๆ ของลูกค้าใหม่ในส่วนของ Customer’s Profile 2. ผู้ใช้งานลบหรือเปลี่ยนแปลงค่า parameter ต่างๆ ได้โดยตรงหรือต้องการ reset ใหม่ด้วยการกดปุ่มคำสั่ง Clear	1.1 ระบบจะเก็บค่าต่างๆ ที่ผู้ใช้งานต้องการนำไปเตรียมไว้สำหรับวิเคราะห์ด้วยแบบจำลองต้นไม้ตัดสินใจที่เลือกไว้ 1.2 ระบบจะเคลียร์ค่าในส่วน Customer’s profile ต่างๆ ให้เป็นค่าว่าง
Exception Conditions:	จำเป็นต้องใส่ค่า parameter ทุกค่าโดยเฉพาะค่าที่มีผลต่อแบบจำลองต้นไม้ตัดสินใจที่เลือกเอาไว้	

ตารางที่ 5.11 แสดง Use Case Description ของ Predict Data

Use Case Name:	9. Predict Data	
Scenario:	วิเคราะห์ข้อมูลของลูกค้ารายใหม่ที่ต้องการ	
Triggering Event:	ผู้ใช้งานกดปุ่มคำสั่ง Predict Data	
Brief Description:	ผู้ใช้งานเลือกคำสั่ง Predict Date เพื่อทำการวิเคราะห์ข้อมูลของลูกค้ารายใหม่ที่ต้องการ	
Actors:	ผู้ใช้งาน	
Related Use Cases:	7. Select Model, 8. Edit Customer's profile	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องใส่ข้อมูลในส่วนของ Customer's Profile	
Postconditions:	ผลลัพธ์จากการวิเคราะห์ข้อมูล	
Flow of Events:	Actor	System
	1. ผู้ใช้งานเลือกปุ่มคำสั่ง Predict Data	1.1 ระบบจะจำค่าต่างๆ ที่เก็บไว้จากผู้ใช้งานต้องการไปวิเคราะห์ด้วยแบบจำลองต้นไม้ตัดสินใจที่เลือกไว้ 1.2 ระบบจะแสดงผลลัพธ์ที่ได้จากการวิเคราะห์ที่ช่อง Predict Result 1.3 ระบบจะแสดงผลลัพธ์ที่ได้ในรูปแบบต้นไม้ตัดสินใจ
Exception Conditions:	-	

ตารางที่ 5.12 แสดง Use Case Description ของ Save Predicted Data

Use Case Name:	10. Save Predicted Data	
Scenario:	เก็บบันทึกผลลัพธ์จากการวิเคราะห์ด้วยแบบจำลองต้นไม้ที่ต้องการ	
Triggering Event:	ผู้ใช้งานกดปุ่มคำสั่ง Save Predicted Data	
Brief Description:	ผู้ใช้งานสามารถเก็บบันทึกผลลัพธ์จากการวิเคราะห์ด้วยแบบจำลองต้นไม้ที่ต้องการไว้และสามารถเรียกดูได้ในภายหลัง	
Actors:	ผู้ใช้งาน	
Related Use Cases:	9. Predict Data	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องมีข้อมูลในส่วนของ Customer's Profile และผลลัพธ์จากการวิเคราะห์ด้วยแบบจำลองต้นไม้	
Postconditions:	ข้อมูล Customer's Profile และผลลัพธ์จากการวิเคราะห์ด้วยแบบจำลองต้นไม้	
Flow of Events:	Actor	System
	1. ผู้ใช้งานคลิกคำสั่ง Save Predicted Data	1.1 ระบบแสดงหน้าจอข้อย่อย Save file
	2. เลือก directory ที่ต้องการเก็บไว้ในหน่วยความจำสำรอง	2.1 ระบบแสดง directory ต่างๆ ของที่ผู้ใช้งานต้องการเก็บข้อมูล
	3. กดคำสั่ง OK เพื่อยืนยันการ Save Predicted Data	3.1 ระบบทำการเก็บบันทึกข้อมูลลงฐานข้อมูลใน table tbresult
Exception Conditions:	-	

ตารางที่ 5.13 แสดง Use Case Description ของ Search Predicted data

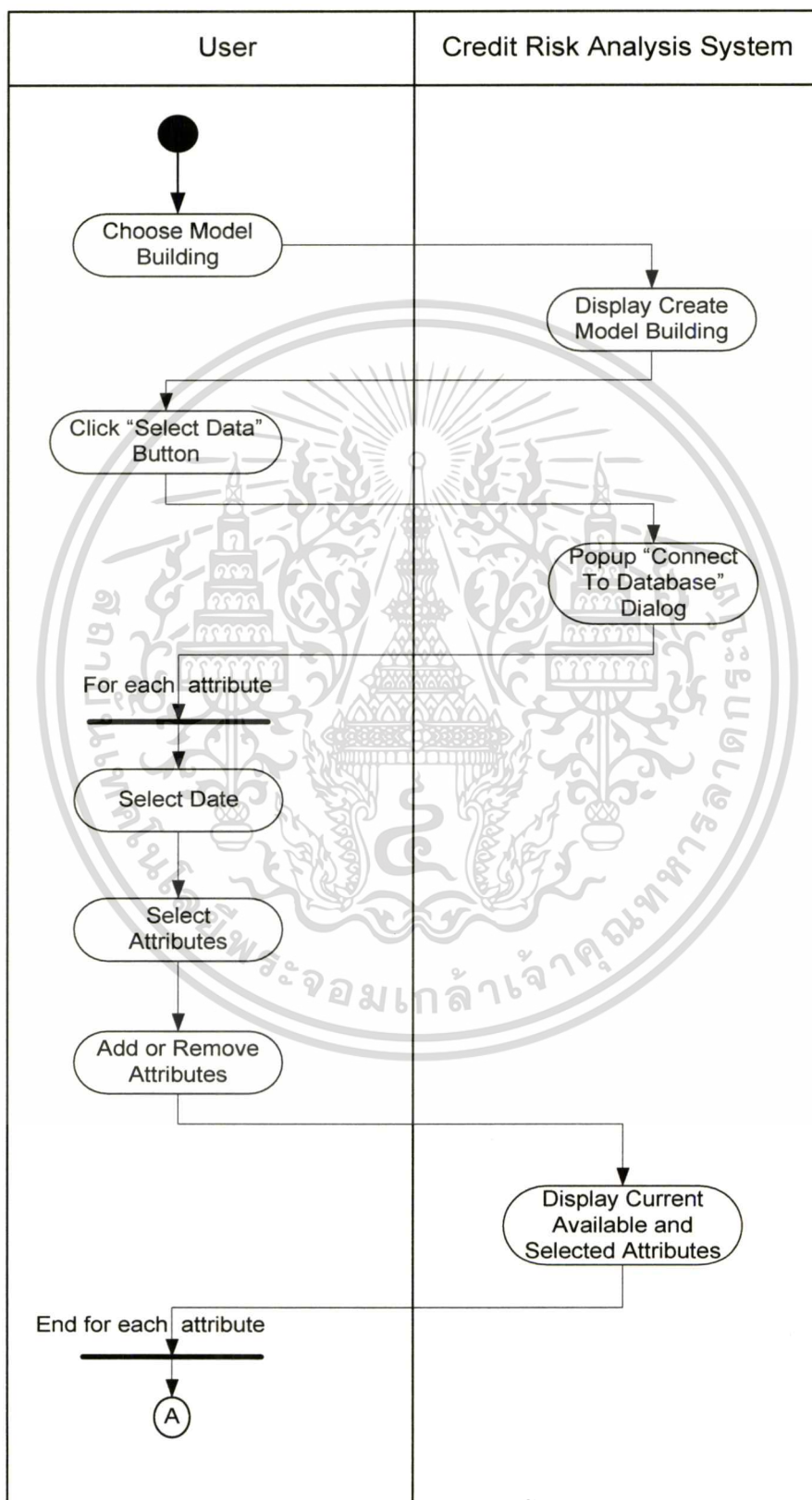
Use Case Name:	11. Search Predicted data	
Scenario:	ค้นหาข้อมูลลูกค้าที่ผ่านการวิเคราะห์ด้วยแบบจำลองต้นไม้แล้ว	
Triggering Event:	ผู้ใช้งานกดปุ่มคำสั่ง Search Predicted data	
Brief Description:	ผู้ใช้งานสามารถค้นหาข้อมูลลูกค้าและผลลัพธ์จากการวิเคราะห์ด้วยแบบจำลองต้นไม้	
Actors:	ผู้ใช้งาน	
Related Use Cases:	9. Predict Data	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	-	
Postconditions:	ข้อมูลของลูกค้าและผลลัพธ์ตามเงื่อนไขที่ผู้ใช้งานต้องการค้นหา	
Flow of Events:	Actor	System
	1. ผู้ใช้งานเลือกค่าแอททริบิวต์ที่ต้องการ 2. ผู้ใช้งานเลือกเครื่องหมายในการเปรียบเทียบ (Logical Funtion) เช่น เครื่องหมายมากกว่า, น้อยกว่า, เท่ากับ เป็นต้น 3. ผู้ใช้งานใส่ค่า Parameter ที่ต้องการ 4. ผู้ใช้งานกดปุ่มคำสั่ง Search เพื่อยืนยันการค้นหาข้อมูล	1.1 ระบบจะแสดงค่าแอททริบิวต์ต่างๆ 2.1 ระบบจะแสดงเครื่องหมายในการเปรียบเทียบ (Logical Funtion) 3.1 ระบบจะเก็บค่า Parameter ที่ผู้ใช้งานใส่ไว้เพื่อนำไปใช้ในการค้นหา 4.1 ระบบจะทำการประมวลผลและแสดงผลลัพธ์ในรูปแบบตารางในส่วนล่างของหน้าจอ
Exception Conditions:	การเลือกค้นหาแต่ละครั้งจะค้นหาได้เพียงแอททริบิวต์เดียวที่สนใจ	

ตารางที่ 5.14 แสดง Use Case Description ของ View Predicted data table

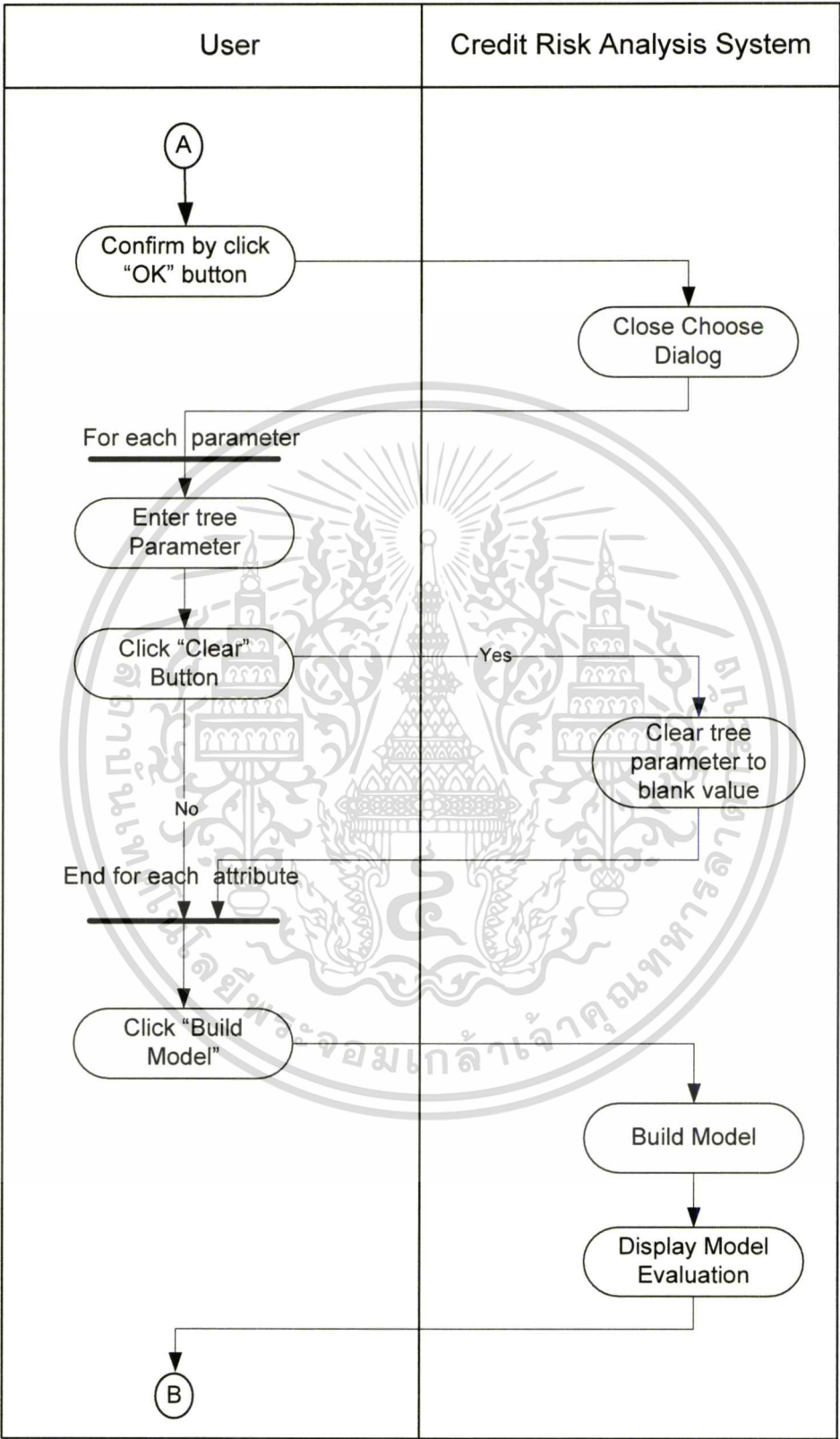
Use Case Name:	12. View Predicted data table	
Scenario:	เลือกดูผลลัพธ์จากการค้นหาจากหน้าต่างแสดงผลลัพธ์	
Triggering Event:	-	
Brief Description:	ผู้ใช้งานสามารถเลื่อนหน้าต่างและย่อขยายขนาดของแต่ละช่องได้ตามต้องการ	
Actors:	ผู้ใช้งาน	
Related Use Cases:	Search Predicted data	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	-	
Postconditions:	-	
Flow of Events:	Actor	System
	1.ผู้ใช้งานเลื่อนหน้าต่างผลลัพธ์, ย่อหรือขยายขนาดของแต่ละช่อง	1.1 ระบบจะแสดงหน้าต่างผลลัพธ์ตามผู้ใช้งานต้องการ
Exception Conditions:	ไม่สามารถ Copy ข้อมูลที่ต้องการไปใช้งานได้เลย	

5.4.3 Activity Diagram ของระบบ

แบบจำลอง Activity Diagram อธิบายการทำงานของระบบแสดงดังรูป

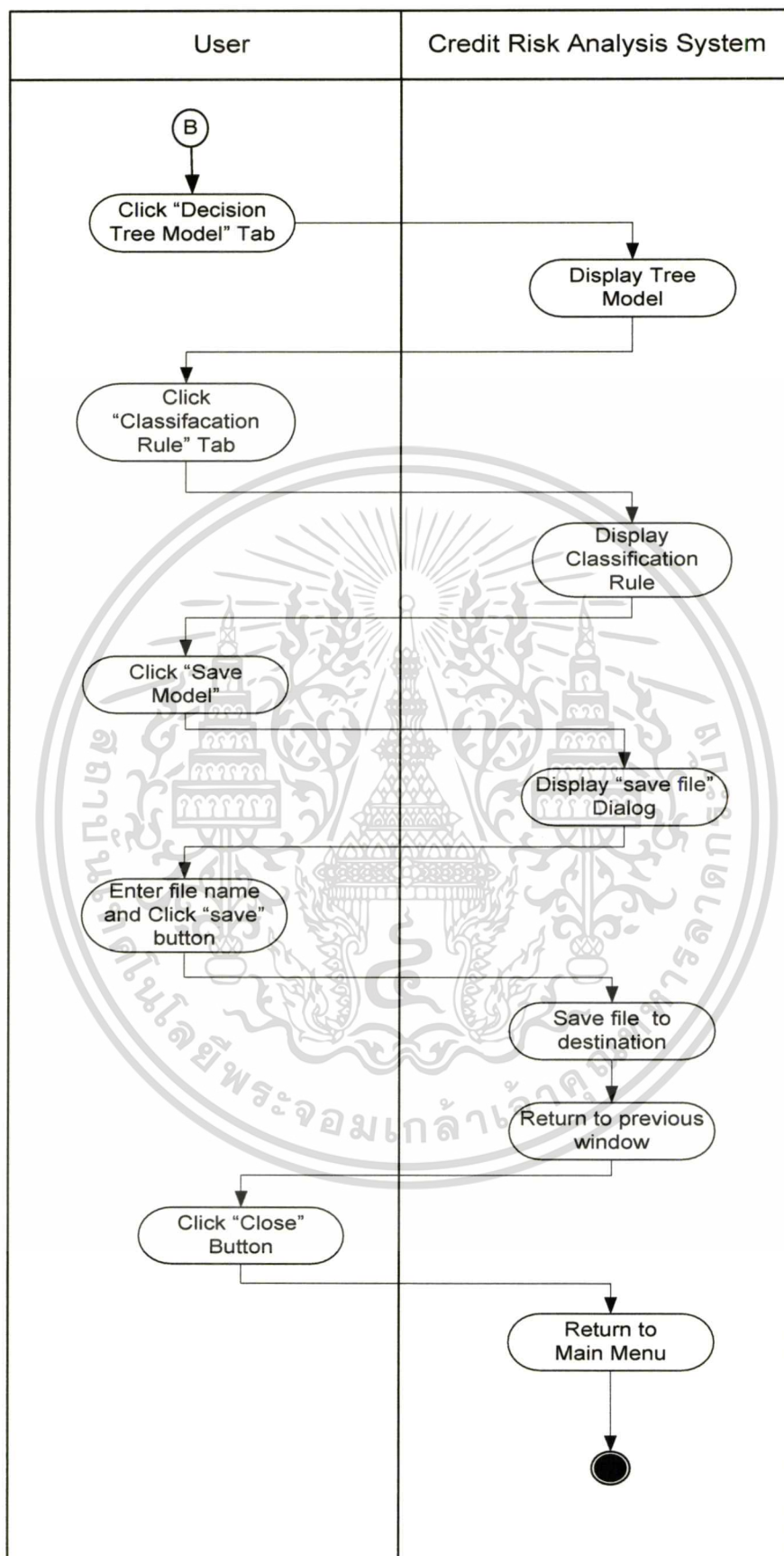


เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
 รูปที่ 5.2 Activity Diagram แสดงการทำงานของระบบ
 ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 5.2 Activity Diagram แสดงการทำงานของระบบ (ต่อ)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

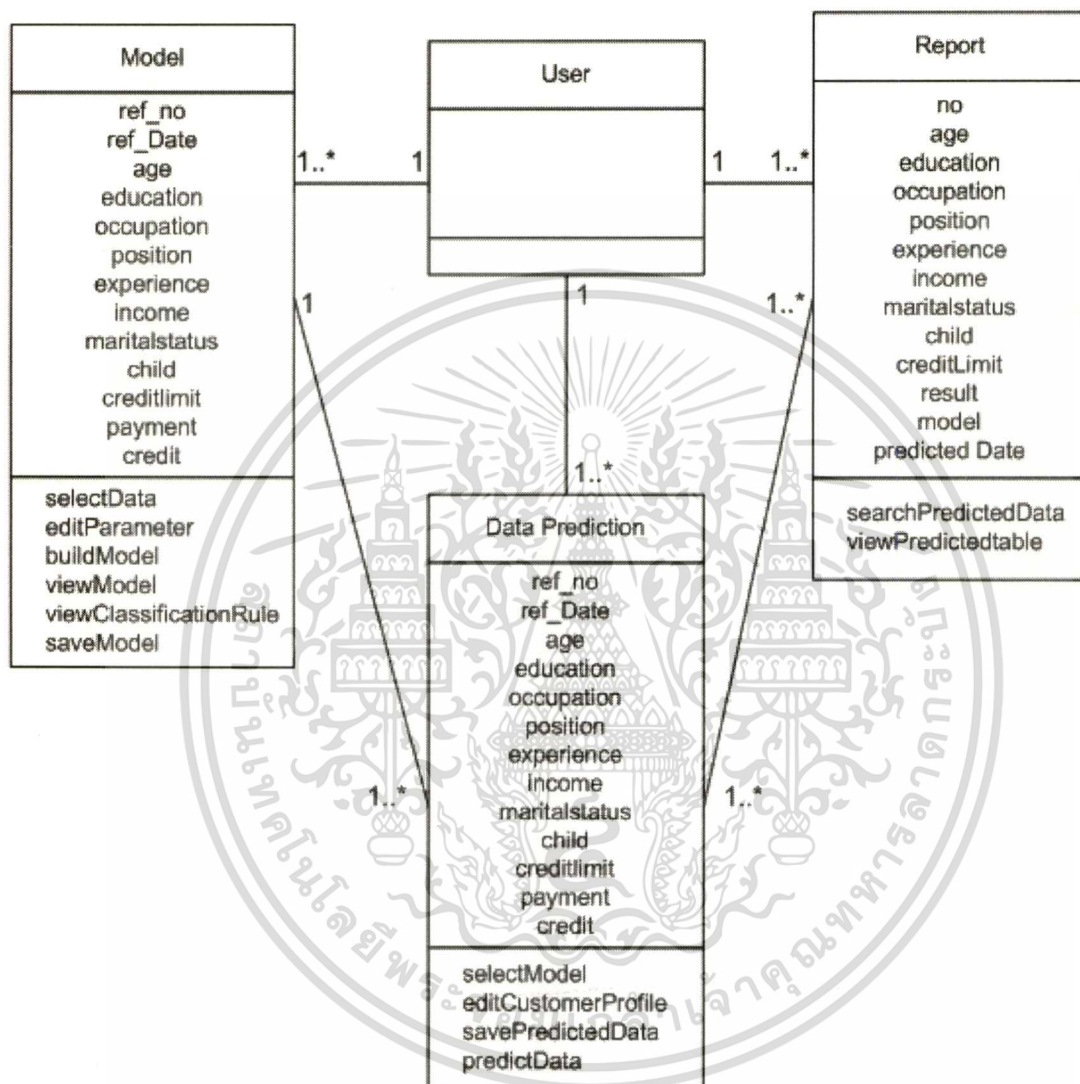


รูปที่ 5.2 Activity Diagram แสดงการทำงานของระบบ (ต่อ)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ตัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.4.4 Class Diagram ของระบบ

Class Diagram แสดงแต่ละ Class ของฟังก์ชันการทำงานและข้อมูลในระบบ และแอททริบิวต์ของแต่ละ Class รวมทั้งความสัมพันธ์ของแต่ละ Class ในระบบดังรูปที่ 5.3



รูปที่ 5.3 แสดง Class Diagram ของระบบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.5 การใช้งานระบบช่วยวิเคราะห์ความเสี่ยงของลูกค้าโทรศัพท์เคลื่อนที่

ในการพัฒนาระบบช่วยวิเคราะห์ความเสี่ยงของลูกค้าโทรศัพท์เคลื่อนที่ด้วยดัชนีชั้นทรินั้นมีการออกแบบให้ระบบมีส่วนการทำงานหลักออกเป็น 3 ส่วนดังนี้

1. การสร้างแบบจำลองพยากรณ์ (Model Building) โดยมีส่วนการทำงานย่อยดังนี้

- การเลือกข้อมูลจากฐานข้อมูลเพื่อนำมาสร้างแบบจำลองพยากรณ์
- การสร้างแบบจำลองพยากรณ์ในรูปแบบต้นไม้ตัดสินใจ
- การสร้างแบบจำลองพยากรณ์ในรูปแบบกฎการตัดสินใจ If-Then
- การบันทึกแบบจำลองเพื่อใช้ในการพยากรณ์ข้อมูลใหม่

2. การพยากรณ์ข้อมูล (Data Prediction) มีส่วนการทำงานย่อยดังนี้

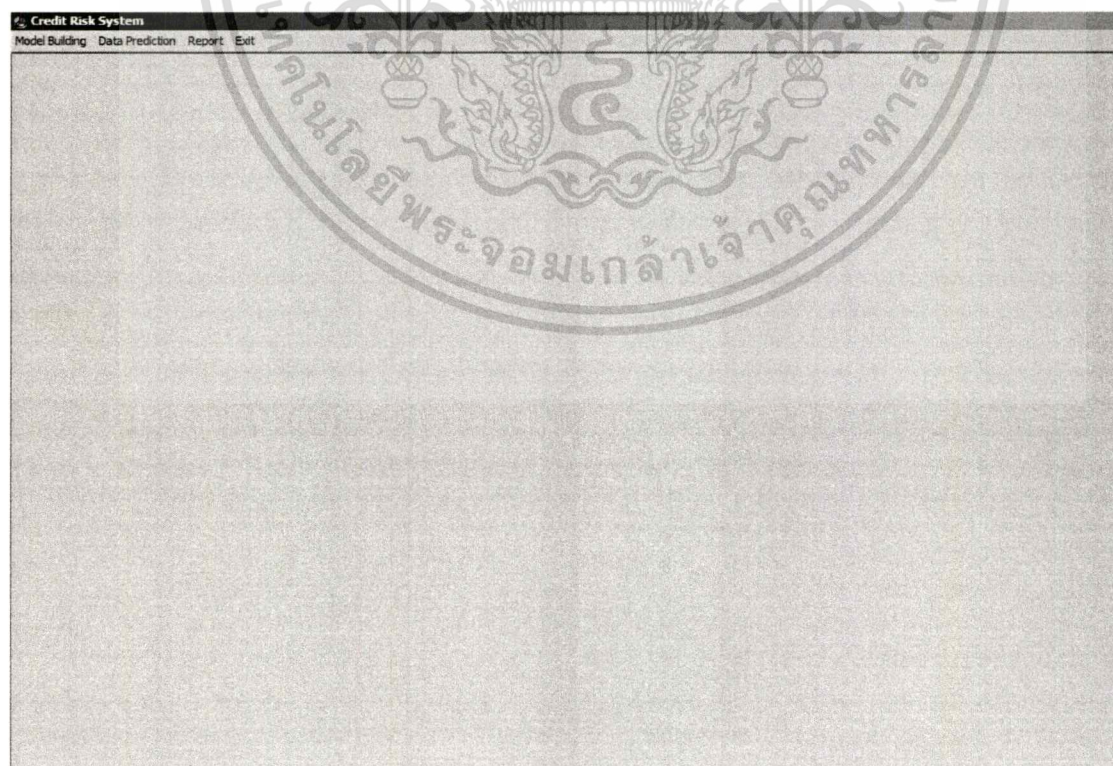
- การเลือกแบบจำลองพยากรณ์ที่ได้บันทึกไว้เพื่อนำมาใช้พยากรณ์
- การพยากรณ์ข้อมูลใหม่
- การบันทึกผลการพยากรณ์

3. การค้นหาและรายงาน (Report) มีส่วนการทำงานย่อยดังนี้

- การค้นหาผลการพยากรณ์และการแสดงผลการพยากรณ์

เมื่อเข้าสู่ระบบนั้น ระบบจะแสดงหน้าจอหลักซึ่งมีเมนูการทำงานหลัก 3 เมนูข้างต้นที่ผ่าน

มาดังรูปที่ 5.4



รูปที่ 5.4 แสดงหน้าจอหลักของโปรแกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

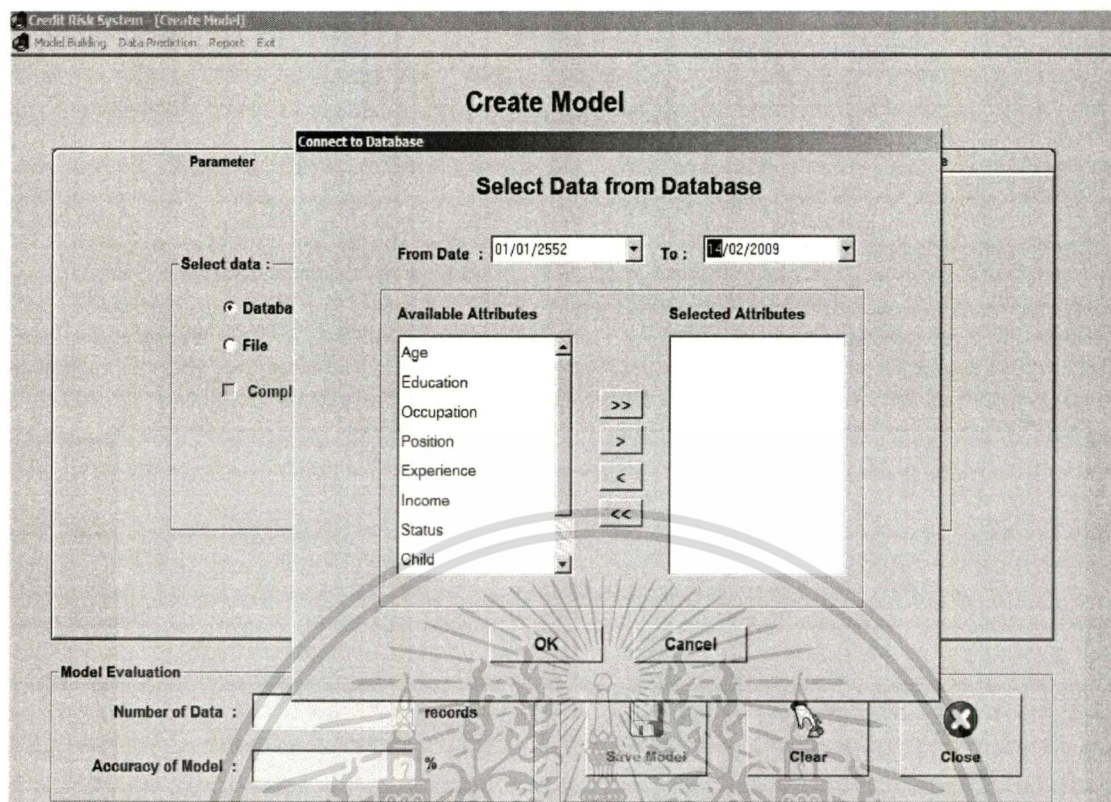
5.4.1 การสร้างแบบจำลองพยากรณ์ (Model Building)

ในขั้นตอนการสร้างแบบจำลองพยากรณ์นั้น เมื่อผู้ใช้งานเข้าสู่เมนู Model Building แล้วผู้ใช้งานจะต้องทำการนำเข้าสู่ข้อมูลจากฐานข้อมูลเพื่อนำมาใช้สร้างแบบจำลองเป็นขั้นตอนแรก โดยผู้ใช้งานสามารถคลิกปุ่ม Select Data ในหน้า Parameter Tab เพื่อเลือก Field ที่เป็นตัวแปรที่จะนำไปใช้ในการวิเคราะห์ตามที่ต้องการ ดังรูปที่ 5.5 และรูปที่ 5.6 ตามลำดับ

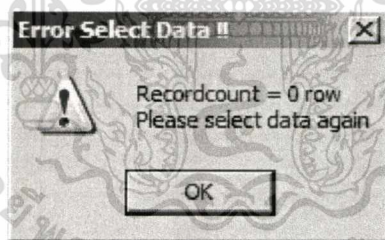
หากการนำเข้าสู่ข้อมูลที่จะใช้วิเคราะห์นั้นไม่สำเร็จนั้นจะมีกล่องข้อความแสดงผิดพลาดให้ผู้ใช้งานทราบดังรูปที่ 5.7 แต่ถ้าหากการนำเข้าสู่ข้อมูลได้สำเร็จนั้นจะไม่แสดงกล่องข้อความผิดพลาดใดๆ และผู้ใช้งานสามารถสังเกตได้จากสถานะของ Check Box ของข้อความ Completed ได้เช่นกัน

รูปที่ 5.5 แสดงเมนูหลัก Create Model

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 5.6 แสดงการเลือกข้อมูลจากฐานข้อมูล



รูปที่ 5.7 แสดงกล่องข้อความผิดพลาดหากไม่พบข้อมูล

เมื่อผู้ใช้งานสามารถนำเข้าข้อมูลและเลือกตัวแปรที่ต้องการใช้ในการวิเคราะห์ข้อมูลได้สำเร็จแล้วนั้น ขั้นตอนต่อไปคือการใส่ค่าพารามิเตอร์เพื่อสร้างแบบจำลองที่ต้องการดังรูปที่ 5.8 โดยต้องใส่ค่าพารามิเตอร์ดังต่อไปนี้

- Maximum Tree Level ใช้กำหนดระดับของแบบจำลองต้นไม้ที่ต้องการสร้าง โดยนับ Root Node เป็นระดับที่ 1 ของแบบจำลองต้นไม้
- Minimum Leaf Node ใช้กำหนดจำนวนของ Leaf Node ของแบบจำลองต้นไม้ที่ต้องการ
- Percentage of Training Data ใช้กำหนดจำนวนของ Training Data ที่ต้องการคิดเป็น

เอกสารนี้เป็นเอกสารต้นฉบับไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Credit Risk System - [Create Model]

Model Building Data Prediction Report Exit

Create Model

Parameter

Decision Tree Model

Classification Rule

Select data :

- ☒ Database
- ☐ File
- ☒ Completed

Select Data

Enter parameter :

Maximum Tree Level :

Minimum Quantity at Leaf Node :

Percentage of Training Data : %

Build Model

Model Evaluation

Number of Data : records

Accuracy of Model : %

Save Model

Clear

Close

รูปที่ 5.8 แสดงตัวอย่างการใส่ค่าพารามิเตอร์เพื่อสร้างแบบจำลองแสดง

และเมื่อผู้ใช้งานใส่ค่าพารามิเตอร์ที่ต้องการสร้างแบบจำลองเสร็จแล้ว ผู้ใช้สามารถกดปุ่ม Build Model เพื่อให้โปรแกรมสร้างแบบจำลองได้ ซึ่งผู้ใช้งานสามารถเลือกดูผลของการสร้างแบบจำลองในรูปแบบต้นไม้โดยกด Tab ชื่อ Decision Tree Model ดังรูปที่ 5.9 หรือแบบกฎการตัดสินใจดังรูปที่ 5.10

Credit Risk System - [Create Model]

Model Building Data Prediction Report Exit

Create Model

Parameter

Decision Tree Model

Classification Rule

N1-Start

- N11 : Education = 5 (P)
 - N111 : Age = 47 (P)
 - N112 : Age <> 47 (P)
- N12 : Education <> 5 (N)
 - N121 : Age = 43 (P)
 - N122 : Age <> 43 (N)

Model Evaluation

Number of Data : records

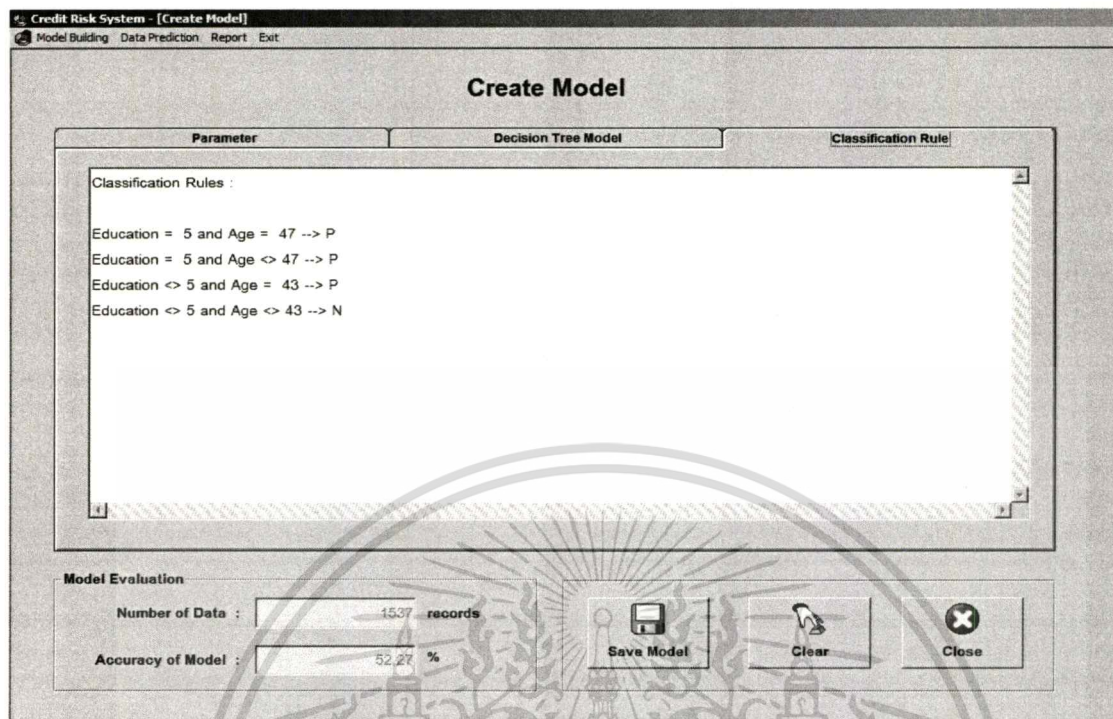
Accuracy of Model : %

Save Model

Clear

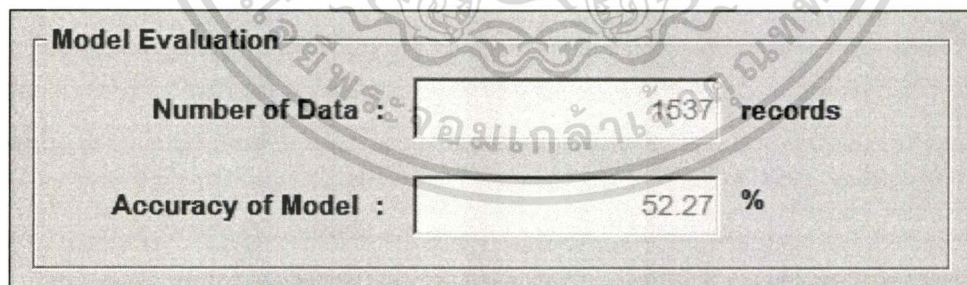
Close

เอกสารนี้เป็นเอกสารที่รูปที่ 5.9 แสดงแบบจำลองที่สร้างในรูปแบบต้นไม้ตัดสินใจนำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



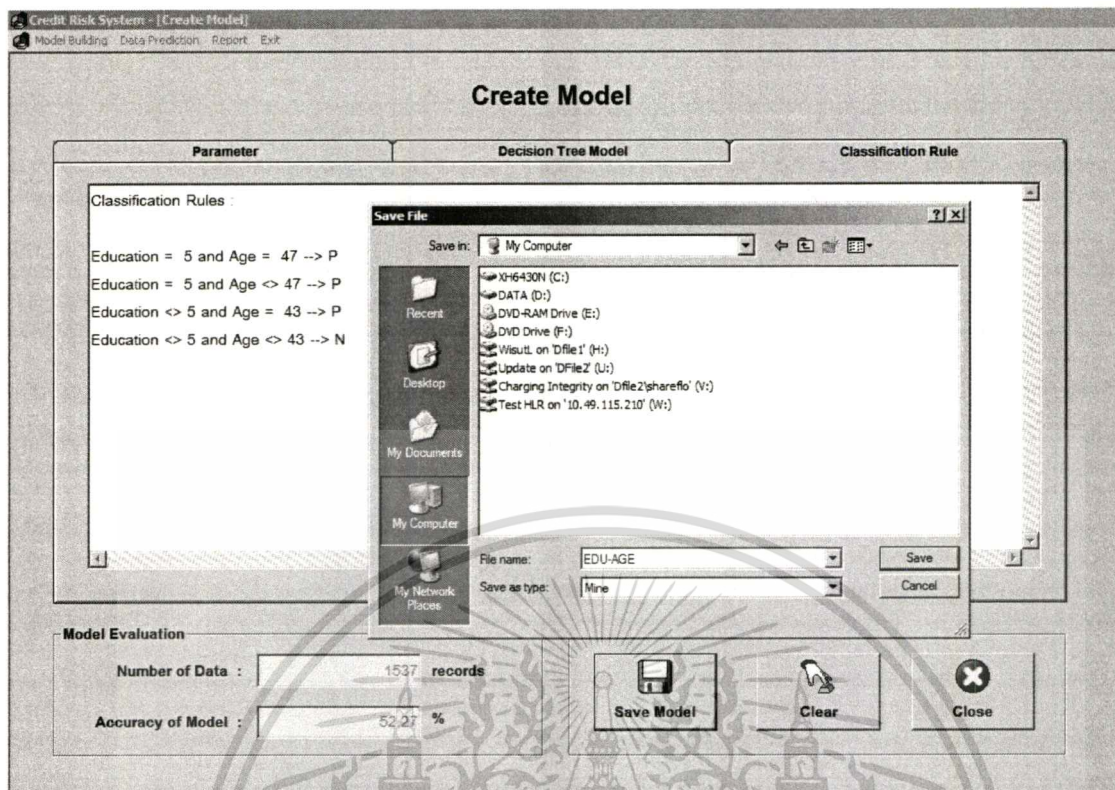
รูปที่ 5.10 แสดงแบบจำลองที่สร้างในรูปกฎการตัดสินใจ if-then

ในการประเมินแบบจำลองที่สร้างขึ้นนั้นผู้ใช้งานสามารถดูได้จากส่วนของ Model Evaluation ด้านล่างซ้ายของหน้าจอ ดังรูปที่ 5.11 โดยข้อมูล Number of Data แสดงจำนวนระเบียบข้อมูลของ Training Data ที่ใช้ และ Accuracy of Model แสดงเปอร์เซ็นต์ของความถูกต้องของแบบจำลอง



รูปที่ 5.11 แสดงข้อมูลของ Training Data และการประเมินแบบจำลอง

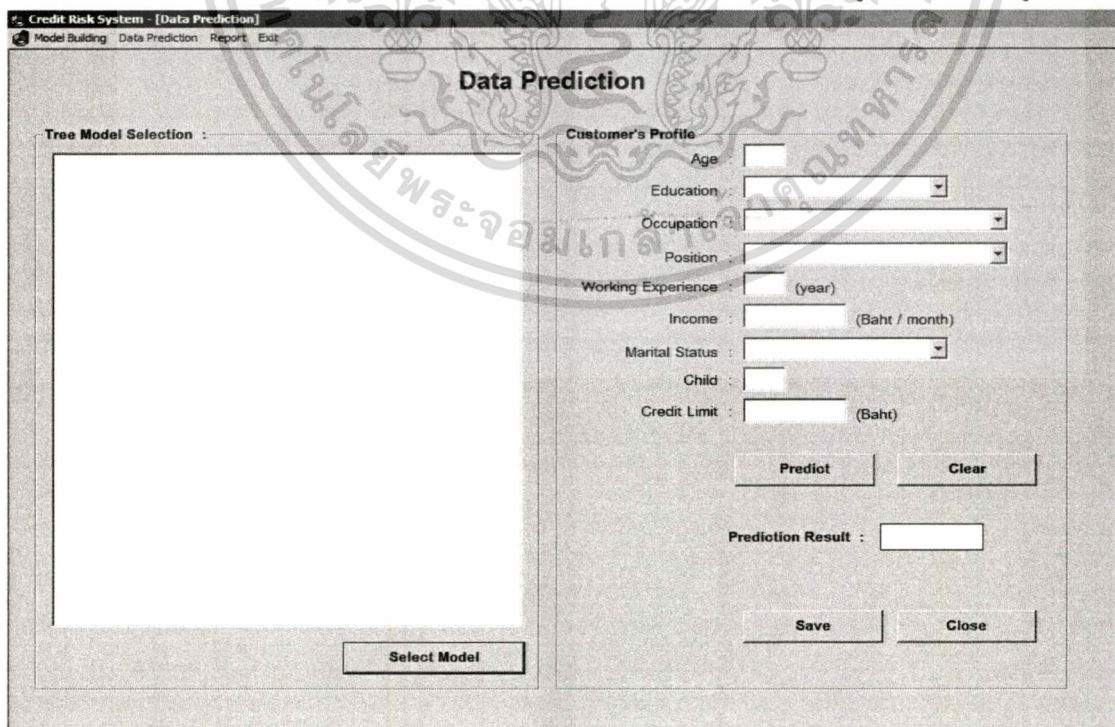
หากผู้ใช้งานต้องการบันทึกแบบจำลองที่สร้างไว้เพื่อใช้วิเคราะห์ข้อมูลต่อไป สามารถกดปุ่ม Save Model เพื่อบันทึกแบบจำลองได้ตามต้องการ ดังรูป 5.12



รูปที่ 5.12 แสดงหน้าจอเพื่อบันทึกแบบจำลองที่สร้างไว้

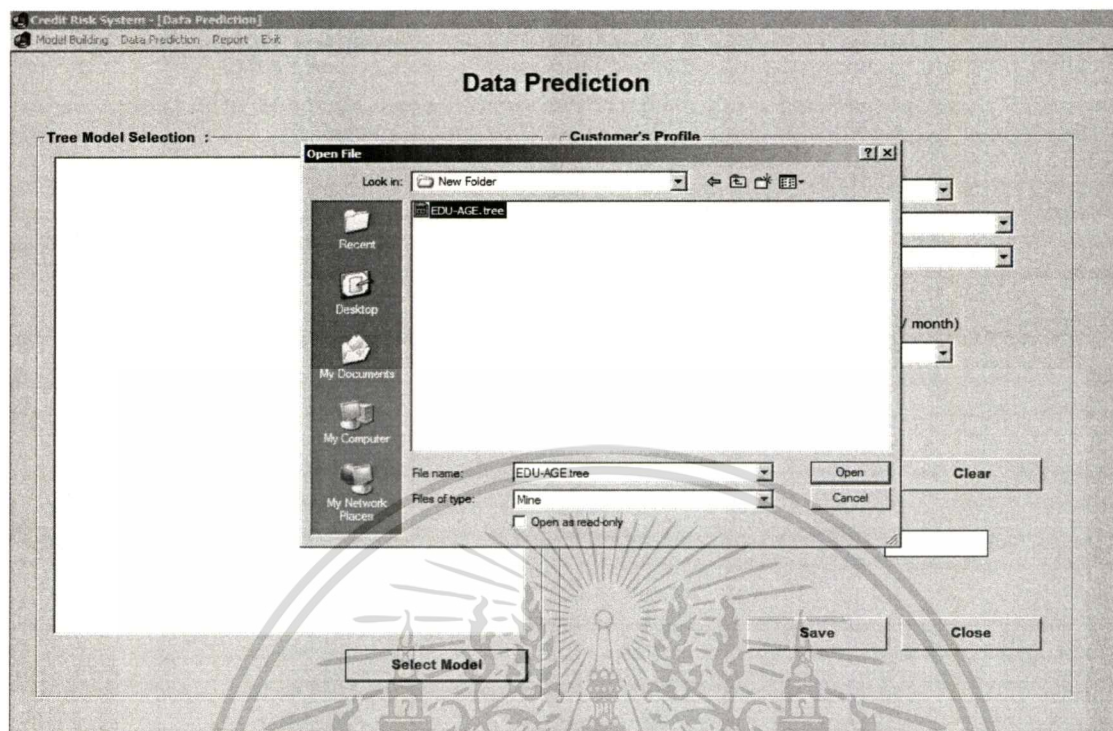
5.4.2 การพยากรณ์ข้อมูล (Data Prediction)

ในขั้นตอนการพยากรณ์ข้อมูลนั้น เมื่อผู้ใช้งานเข้าสู่เมนู Data Prediction แล้วผู้ใช้งานสามารถที่จะทำการเลือกแบบจำลองที่ต้องการซึ่งได้บันทึกไว้ก่อนหน้านี้ดังรูปที่ 5.13 และรูปที่ 5.14

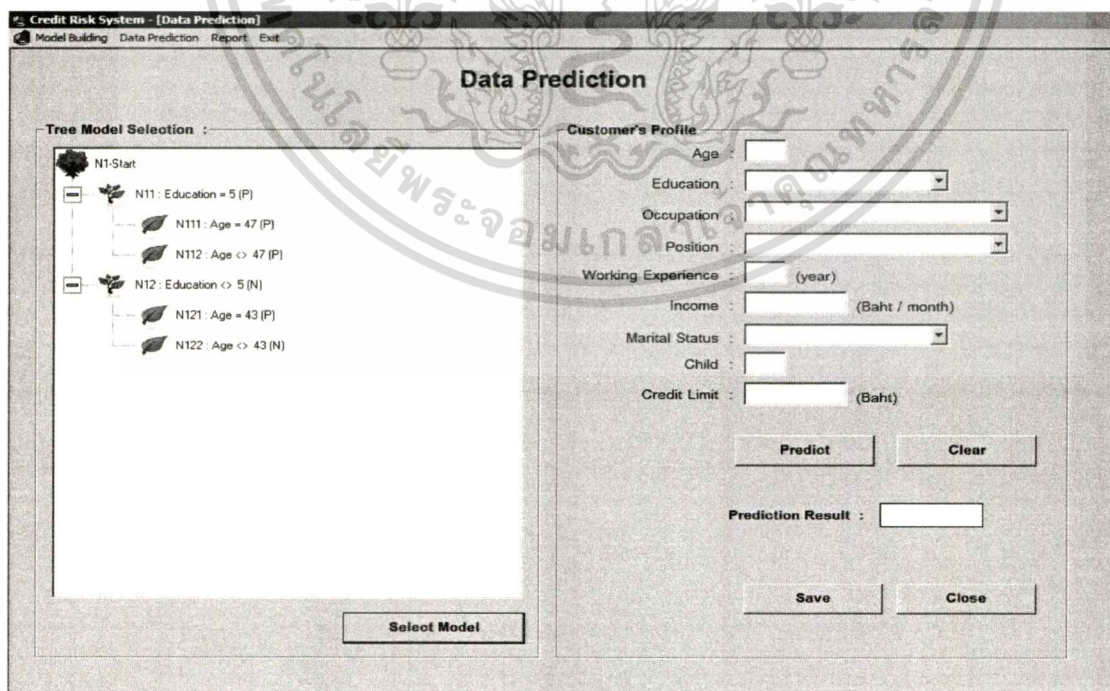


รูปที่ 5.13 แสดงเมนูหลัก Data Prediction

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 5.14 แสดงหน้าจอเพื่อเลือกแบบจำลองที่ต้องการใช้พยากรณ์
เมื่อผู้ใช้เลือกแบบจำลองที่ต้องการได้แล้ว ส่วนของหน้าจอ Tree Model จะแสดง
แบบจำลองในรูปแบบของต้นไม้ตัดสินใจดังรูปที่ 5.15



รูปที่ 5.15 แสดงแบบจำลองที่ถูกเลือกใช้พยากรณ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ผู้ใช้งานสามารถพยากรณ์ข้อมูลของลูกค้าใหม่ที่เราต้องการพยากรณ์ด้วยแบบจำลองที่เลือกมาได้โดยใส่ข้อมูลของลูกค้าในส่วน Customer's Profile ดังรูปที่ 5.16

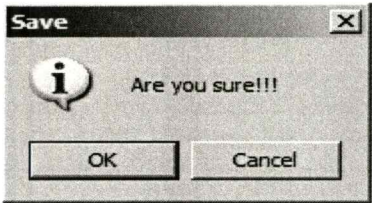
รูปที่ 5.16 แสดงการใส่ข้อมูลของลูกค้าที่ต้องการพยากรณ์

เมื่อผู้ใช้งานใส่ค่าข้อมูลของลูกค้าใหม่และทำการกดปุ่ม Predict เพื่อพยากรณ์ลูกค้ารายใหม่ตามแบบจำลองแล้ว ผู้ใช้งานสามารถดูผลของการพยากรณ์จากข้อมูล Prediction Result หรือหน้าต่างในส่วนของ Tree Model Selection ดังรูปที่ 5.17

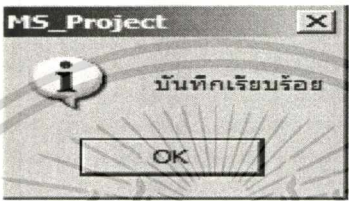
รูปที่ 5.17 แสดงผลการพยากรณ์ของลูกค้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เมื่อทราบผลการพยากรณ์แล้วผู้ใช้งานสามารถบันทึกผลของการพยากรณ์โดยการกดปุ่ม Save ซึ่งข้อมูลของลูกค้าและผลการพยากรณ์จะถูกบันทึกลงในฐานข้อมูลต่อไป



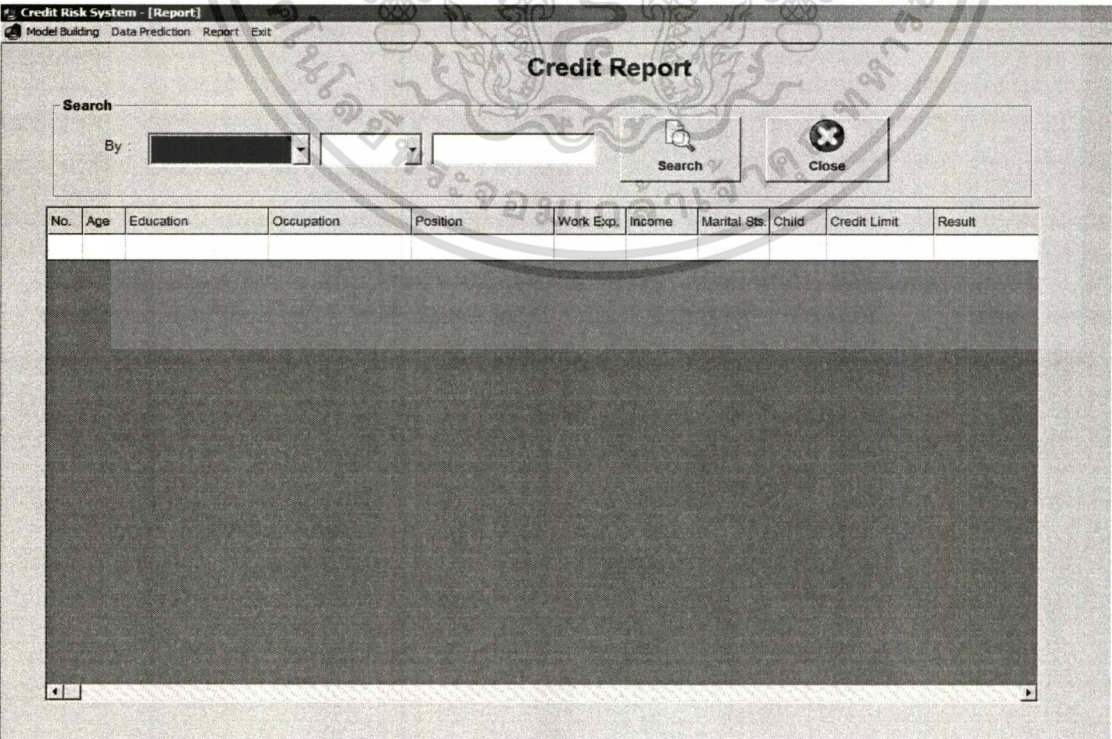
รูปที่ 5.18 แสดงกล่องข้อความเพื่อยืนยันการบันทึกผล



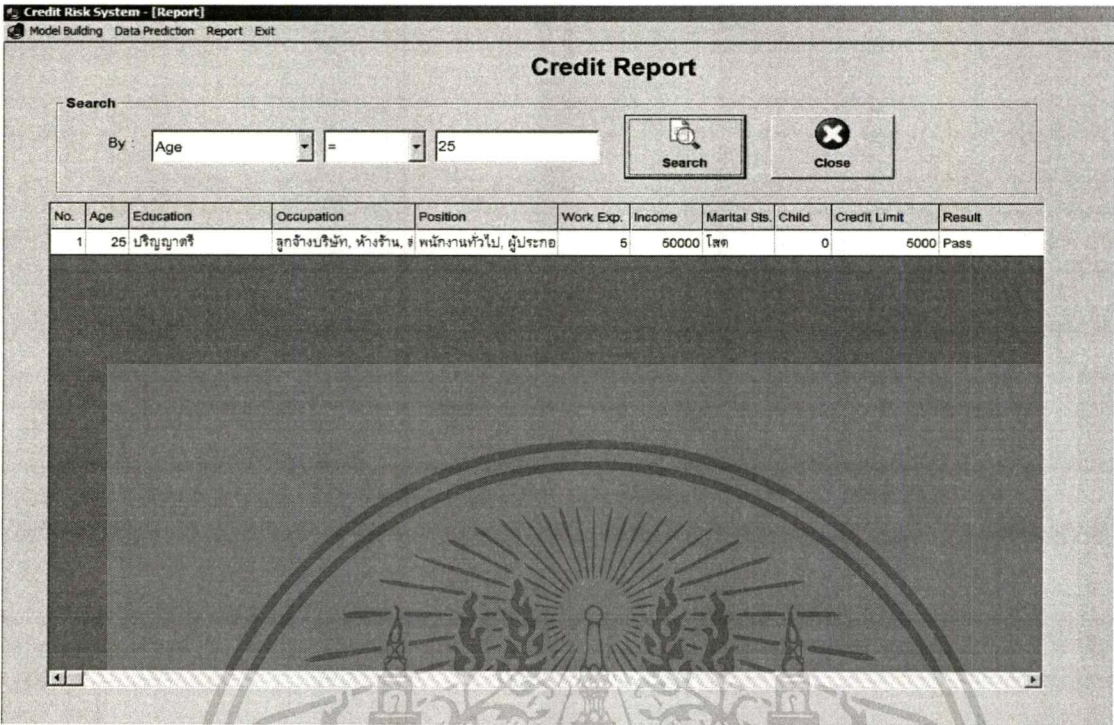
รูปที่ 5.19 แสดงกล่องข้อความเมื่อบันทึกผลสำเร็จ

5.4.3 การค้นหาข้อมูลและรายงาน (Credit Report)

ในขั้นตอนการค้นหาข้อมูลและรายงานนั้น ผู้ใช้งานสามารถเข้าสู่เมนูหลัก Report ดังรูปที่ 5.20 แล้วผู้ใช้งานสามารถที่จะเลือกค้นหาข้อมูลที่เก็บบันทึกไว้ตามตัวแปรที่ใช้ในการวิเคราะห์ หรือตามแบบจำลองที่ใช้งานดังรูปที่ 5.21



เอกสารนี้เป็นเอกสารที่สงวนไว้รูปที่ 5.20 แสดงเมนูหลัก Credit Report
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 5.23 แสดงผลลัพธ์ที่ต้องการ

จากที่ได้อธิบายการทำงานของระบบข้างต้น ซึ่งประกอบด้วยขั้นตอนการเลือกข้อมูลในการสร้างแบบจำลอง การพยากรณ์ข้อมูลตามแบบจำลองที่สร้างขึ้น และการค้นหาข้อมูลผลการพยากรณ์ของลูกค้าโทรศัพท์เคลื่อนที่ไปแล้วนั้น ในบทถัดไปจะกล่าวถึงผลสรุปของการศึกษาและการพัฒนาระบบ รวมทั้งผลสรุปของการดำเนินงานและข้อเสนอแนะต่างๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 6

บทสรุปและข้อเสนอแนะ

ผลจากการศึกษาและพัฒนาระบบช่วยวิเคราะห์ความเสี่ยงของลูกค้าโทรศัพท์เคลื่อนที่เพื่อช่วยวิเคราะห์ความเสี่ยงของเครดิตของลูกค้ารายใหม่ที่จะเข้ามาในระบบนั้นได้ข้อสรุปดังต่อไปนี้

6.1 สรุปผล

จากการศึกษาและนำอัลกอริทึม SLIQ เพื่อใช้ในการทำค้ำไ่มนึ่งเพื่อช่วยวิเคราะห์ความเสี่ยงของลูกค้าโทรศัพท์เคลื่อนที่ด้วยดิซันทรินั้น พบว่าระบบสามารถช่วยวิเคราะห์ข้อมูลได้ตามวัตถุประสงค์ของโครงการ คือ

1. ระบบสามารถช่วยในการวิเคราะห์ข้อมูลได้ตรงตามหลักการของอัลกอริทึม SLIQ
2. ระบบสามารถแสดงผลการวิเคราะห์ในรูปแบบของต้นไม้ตัดสินใจ หรือกฎการตัดสินใจ If-Then ซึ่งผู้ใช้งานสามารถเข้าใจได้ง่าย
3. ระบบสามารถช่วยลดเวลาในการวิเคราะห์ของผู้ใช้งาน
4. ระบบสามารถช่วยเพิ่มความถูกต้องของการวิเคราะห์ข้อมูล แม้ว่าผู้ใช้งานจะไม่มีประสบการณ์ในการวิเคราะห์ข้อมูลความเสี่ยง
5. ระบบสามารถจะนำไปประยุกต์ใช้กับการวิเคราะห์ข้อมูลด้านอื่นๆ ได้อีกโดยคงใช้หลักการของอัลกอริทึม SLIQ และการนำเสนอในรูปแบบต้นไม้ตัดสินใจ

6.2 ประโยชน์ที่ได้รับ

การศึกษาการทำค้ำไ่มนึ่งเพื่อช่วยวิเคราะห์ความเสี่ยงของลูกค้าโทรศัพท์เคลื่อนที่ด้วยดิซันทรินโดยใช้อัลกอริทึม SLIQ นี้ทำให้ได้รับประโยชน์ดังนี้

1. ทำให้มีความเข้าใจและประโยชน์ของการทำค้ำไ่มนึ่งมากขึ้น และสามารถนำทฤษฎีการทำค้ำไ่มนึ่งมาใช้ประโยชน์ได้จริง
2. ทำให้ได้ศึกษาและเข้าใจถึงหลักการของอัลกอริทึม SLIQ
3. ทำให้ทราบถึงวิธีการจัดการและการบริหารโครงการ ซึ่งสามารถนำไปใช้ในการทำงานได้จริง
4. สามารถนำระบบที่พัฒนาไปใช้ประโยชน์ในองค์กรเพื่อช่วยวิเคราะห์ความเสี่ยงได้จริง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

6.3 ข้อเสนอแนะและแนวทางในการพัฒนาโครงการเพิ่มเติม

1. ข้อมูลที่นำมาศึกษาและวิเคราะห์นั้นอาจจะมีจำนวนน้อยเมื่อเทียบกับข้อมูลทั้งหมดที่มีในองค์กร ซึ่งอาจจะมีผลต่อความถูกต้องในการวิเคราะห์ข้อมูลเช่นกัน
2. ข้อมูลที่นำมาวิเคราะห์นั้นอาจจะมีความสัมพันธ์กับข้อมูลอื่นๆ ซึ่งผู้พัฒนายังไม่ได้นำมาวิเคราะห์ในการศึกษาครั้งนี้ด้วย เช่น โปรโมชันที่ถูกค้าเลือกใช้ เป็นต้น ซึ่งข้อมูลดังกล่าวอาจจะมีผลต่อความถูกต้องในการวิเคราะห์ข้อมูลเช่นกัน ถ้าหากต้องการให้ผลการวิเคราะห์มีความถูกต้องมากกว่านี้จะต้องศึกษาเพิ่มเติมว่าข้อมูลใดที่มีความสัมพันธ์และมีผลต่อความเสี่ยงของเครดิตลูกค้า
3. จากการศึกษาการทำค้ำไ่มิ่งนั้น พบว่ามีทฤษฎีและหลักการอีกมากที่สามารถนำมาใช้ในการวิเคราะห์ข้อมูล ซึ่งเป็นไปได้ว่าการทำค้ำไ่มิ่งด้วยอัลกอริทึมอื่นนั้นอาจจะเหมาะสมที่จะใช้วิเคราะห์ข้อมูลลูกค้าโทรศัพท์เคลื่อนที่มากกว่าอัลกอริทึม SLIQ ดังนั้นควรจะมีการศึกษาอัลกอริทึมอื่นๆ เพิ่มเติมด้วย



บรรณานุกรม

- Berry, Michael J.A. and Linoff, Gordon S. 2004. **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management**. Indiana, USA: Wiley Publishing, Inc.
- Berson, Alex, et al. 1999. **Building Data Mining Application for CRM**. Montreal, Canada: McGraw-Hill Companies.
- Han, Jiawei, and Kamber, Micheline. 2000. **Data Mining: Concepts and Techniques**. New York, USA: Morgan Kaufmann.
- M. Mehta, R. Agrawal, and J. Rissanen. 1996. **SLIQ: A fast scalable classifier for data mining**. Avignon, France: In Proc. Int.
- Radosavljevic, Vladan. 2007. **SLIQ**. [Online]. Available: <http://www.cis.temple.edu/~vasilis/Courses/CIS664/Papers/Vladan-SLIQ.ppt>
- Weiss, S. M., and Indurkha, N. 1997. **Predictive data mining: A practical guide**. New York, USA: Morgan-Kaufman.

ประวัติผู้เขียน

ชื่อผู้เขียน

นายวิสูตร ลีลาชุดิพงศ์

วุฒิการศึกษาระดับมัธยมศึกษา

โรงเรียนวิเชียรมาตุ จ.ตรัง

วุฒิการศึกษาระดับปริญญาตรี

วท.บ. (วิทยาการคอมพิวเตอร์)

คณะวิทยาศาสตร์

มหาวิทยาลัยรามคำแหง

ปีที่สำเร็จการศึกษา

2542



เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้