การพัฒนาระบบดาต้าไมนิ่งในการวิเคราะห์หาสาเหตุการยกเลิกบริการโทรศัพท์เคลื่อนที่ โดยใช้ดิซิชั่นทรี

System Development of Data Mining for Churn Management in Mobile Service Using Decision Tree

โดย

นางสาวอุรชา โพธิ์นิ่มแดง

รหัส 43067179

อาจารย์ที่ปรึกษา

ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน (System Development Project)
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
ภาคเรียนที่ 2 ปีการศึกษา 2545
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การพัฒนาระบบคาต้าไมนิ่งในการวิเคราะห์หาสาเหตุการยกเลิกบริการโทรศัพท์เคลื่อนที่โดยใช้ดิซิชั่นทรี
นักศึกษา	นางสาว อุรชา โพธิ์นิ่มแดง
อาจารย์ที่ปรึกษา	ผศ.ดร. วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2545

บทคัดย่อ

ปัจจุบันปัญหาหนึ่งที่ผู้ให้บริการโทรศัพท์เคลื่อนที่กำลังประสบอยู่ คือ ความไม่คงที่ของจำนวนลูกค้าในส่วนแบ่งการตลาดที่มีอยู่ เนื่องจากปัญหาการคืนเลขหมาย และการเปลี่ยนระบบในการใช้บริการยังมีอย่างต่อเนื่อง ส่งผลให้ผู้ประกอบการแต่ละรายต้องหากลยุทธ์ที่ดีที่สุดออกมาใช้เพื่อสามารถรักษาฐานลูกค้าไว้ให้ได้มากที่สุด ดังนั้นโครงงานนี้ได้ทำการศึกษาเทคนิค และพัฒนาระบบการวิเคราะห์หาสาเหตุ และทำนายกลุ่มลูกค้าที่มีโอกาสยกเลิก หรือเปลี่ยนระบบในการใช้บริการ โดยใช้ดิซิชั่นทรีซึ่งเป็นเทคนิคหนึ่งในคาต้าไมนิ่งในการแบ่งกลุ่มลูกค้าที่จะยกเลิก หรือเปลี่ยนระบบในการใช้บริการ และลูกค้าที่จะยังคงใช้บริการต่อไป ซึ่งดิซิชั่นทรีที่ได้สามารถเป็นประโยชน์ต่อผู้ประกอบการในการนำไปใช้ในการวางแผนกลยุทธ์เพื่อรักษาฐานลูกค้าต่อไป

Title	System Development of Data Mining for Churn Management in Mobile Service Using Decision Tree
Student	Ms. Uracha Phonimdang
Advisor	Dr. Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2002

ABSTRACT

Nowadays, Mobile Service Provider faces with a problem that inconsistent number of customer in market share. Because of the mobile phone number cancellation and customer churn problems, so each provider find the best strategy to keep customer base. This project have studied and developed the system to analyze why customers churn and which customers are most likely to churn in future by using decision tree that is a technique in data mining. This technique brings useful modeling analytic for provider make strategy to keep customer base.

กิตติกรรมประกาศ

การจัดทำโครงการพัฒนาระบบการวิเคราะห์หาสาเหตุการยกเลิกการใช้บริการโทรศัพท์เคลื่อนที่โดยใช้ดิซิชั่นทรีครั้งนี้ สำเร็จลุล่วงไปได้ด้วยดี เนื่องจาก คำแนะนำ และความช่วยเหลือจากบุคคลต่างๆ ดังต่อไปนี้

บิดา และมารดา ที่ให้การสนับสนุนการศึกษา ช่วยเหลือ และให้กำลังใจในการฝ่าฟันอุปสรรคต่างๆ จนสำเร็จการศึกษา

ท่าน ผศ.ดร. วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาที่ให้คำปรึกษา และแนะนำแนวทางในการศึกษา และพัฒนาโครงการ ตั้งแต่เริ่มต้น จนกระทั่งสำเร็จตามเป้าหมายด้วยดี

เพื่อนๆ ที่ให้คำแนะนำ และความช่วยเหลือ พร้อมทั้งให้กำลังใจในการพัฒนาโครงงานนี้มาโดยตลอด ได้แก่

คุณ วริน มอญเจริญ
คุณ ฉัตร วัฒนศิริเกียรติ
คุณ ดุจดาว บูรณะพาณิชย์กิจ
คุณ สงกรานต์ ศรีปัญญา
และเพื่อนๆ ทุกคน

จึงใคร่ขอขอบคุณบุคคลดังกล่าวมา ณ โอกาสนี้

อุรชา โพธิ์นิ่มแดง
16 กุมภาพันธ์ 2546

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญภาพ

หน้า

ภาพที่

สารบัญภาพ (ต่อ)

หน้า

ภาพที่

บทที่ 1
บทนำ

1.1 บทนำ

การเจริญเติบโตที่เพิ่มขึ้นในธุรกิจโทรศัพท์เคลื่อนที่ ทำให้การแข่งขันเพื่อให้ได้ส่วนแบ่งทางการตลาดเพิ่มสูงขึ้น ในปัจจุบันผู้ให้บริการในธุรกิจการสื่อสารเกิดขึ้นใหม่มากมาย ลูกค้าจึงมีโอกาสในการใช้บริการมากขึ้น ผู้ให้บริการแต่ละรายต่างก็มีเป้าหมายเดียวกันคือรักษาลูกค้าเดิมและเพิ่มฐานลูกค้าใหม่ และพยายามนำกลยุทธ์ทางการตลาดใหม่ๆ มาใช้เพื่อสร้างแรงจูงใจให้ลูกค้าหันมาเลือกใช้บริการมากขึ้น ดังนั้นการวางแผนกลยุทธ์ที่ดีจะช่วยเพิ่มความได้เปรียบทางการตลาด การวางแผนกลยุทธ์ที่ดีต้องอาศัยปัจจัยหลายอย่าง เช่น บุคลากรที่มีความสามารถ ข้อมูล หรือเทคนิคการวิเคราะห์ข้อมูลที่มีประสิทธิภาพและทันสมัย เป็นต้น ปัจจัยหนึ่งที่มีความสำคัญมากคือข้อมูล ซึ่งเกิดจากการทำธุรกรรมในแต่ละวันของลูกค้า และองค์กร นอกจากจะนำมาใช้ในการดำเนินงานทางธุรกิจให้ประสบผลสำเร็จสมบูรณ์แล้ว ยังสามารถนำมาใช้ให้เกิดประโยชน์ในการวิเคราะห์เพื่อหาโอกาสทางการตลาดใหม่ๆได้ แต่องค์กรส่วนใหญ่ยังมีการใช้ประโยชน์จากข้อมูลได้น้อยมาก

ด้วยเทคนิคต่างๆ ของดาต้าไมนิ่งจะช่วยในการค้นหาข้อมูลความรู้ใหม่ที่ถูกต้อง และนำไปใช้ได้จากฐานข้อมูลขนาดใหญ่ เช่นคลังข้อมูล (Data Warehouse) เป็นต้น ดาต้าไมนิ่งจึงเป็นเทคนิคหนึ่งที่เริ่มได้รับความนิยมในการนำมาประยุกต์ใช้ในธุรกิจการสื่อสาร เพื่อนำมาพัฒนาองค์กร หรือเป็นแนวทางในการดำเนินธุรกิจในรูปแบบใหม่ๆ เช่น ทำนายว่าลูกค้าคนใดมีโอกาสที่จะยกเลิกบริการ หรือเปลี่ยนไปใช้บริการกับผู้ให้บริการรายอื่น, เจาะกลุ่มลูกค้าเป้าหมายที่ชอบซื้อสินค้าหรือบริการหนึ่งคู่กับอีกสินค้า หรือบริการหนึ่ง, ออกแบบสินค้าและบริการตรงตามที่กลุ่มเป้าหมายต้องการ, ทำนายรูปแบบ และระยะเวลาในการใช้โทรศัพท์เพื่อตรวจจับการใช้ที่ผิดปกติ หรือการโกง, ตรวจหาความเสี่ยงในการใช้บริการเกินวงเกินที่จำกัด เป็นต้น เพื่อปกป้องและเพิ่มส่วนแบ่งทางการตลาดต่อไป

1.2 วัตถุประสงค์ในการพัฒนาระบบ

1. ศึกษาเทคนิคทางดาต้าไมนิ่ง โดยเลือกใช้โมเดลในการทำนายที่เป็นการแบ่งกลุ่มแบบดิซิชั่นทรี

2. ใช้ SLIQ อัลกอริทึมซึ่งเป็นเทคนิคทางดาต้าไมนิ่ง มาวิเคราะห์หาสาเหตุการยกเลิกบริการ หรือเปลี่ยนไปใช้บริการโทรศัพท์เคลื่อนที่กับผู้ให้บริการรายอื่น
3. ระบบที่พัฒนาสามารถนำไปเป็นประโยชน์ในการสนับสนุนการวิเคราะห์หาสาเหตุการยกเลิกบริการ หรือเปลี่ยนไปใช้บริการโทรศัพท์เคลื่อนที่กับผู้ให้บริการรายอื่น

1.3 ขอบเขตของการพัฒนาระบบ

โครงการพัฒนาระบบงานนี้ จะนำข้อมูลลูกค้าในธุรกิจโทรศัพท์เคลื่อนที่ที่ได้จากขั้นตอนต่างๆ ของดาต้าไมนิ่ง คือการเลือกข้อมูล, การเตรียมข้อมูลก่อนประมวลผล และการแปลงข้อมูลแล้ว มาทำการสร้างต้นไม้(Decision Tree) ซึ่งจะเป็นโมเดลในการทำนายหาสาเหตุ และกลุ่มลูกค้าที่มีโอกาสเปลี่ยนระบบในการใช้บริการโทรศัพท์เคลื่อนที่ ในอนาคต

1.4 ขั้นตอนการดำเนินงาน

1. กำหนดวัตถุประสงค์ในการวิเคราะห์หาสาเหตุการยกเลิกบริการในธุรกิจโทรศัพท์เคลื่อนที่
2. ศึกษาขั้นตอน และวิธีการทางดาต้าไมนิ่ง โดยเลือก SLIQ Classification เพื่อสร้างโมเดลในการทำนายหาสาเหตุของการยกเลิกบริการ
3. เลือกแหล่งข้อมูลที่จะนำมาใช้ในการสร้างโมเดล และใช้ในการทำนาย โดยเลือกใช้ข้อมูลของลูกค้าที่ใช้บริการอยู่ในปัจจุบัน และข้อมูลของลูกค้าที่ยกเลิกบริการไปแล้ว
4. เตรียม และแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมในการนำมาใช้กับอัลกอริทึม
5. ออกแบบ และพัฒนาระบบ โดยใช้อัลกอริทึม SLIQ Classification

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการศึกษาเทคนิคของดาต้าไมนิ่งโดยนำมาประยุกต์ใช้กับธุรกิจโทรศัพท์เคลื่อนที่ในโครงงานนี้ ทำให้ได้รับความรู้ และความเข้าใจในขั้นตอนต่างๆ ของดาต้าไมนิ่ง และเทคนิคการทำนายโดยการแบ่งกลุ่มแบบดิซิชั่นทรี และสามารถใช้ระบบในการสร้างโมเดลดิซิชั่นทรี ซึ่งสามารถนำไปเป็นแนวทางให้กับผู้ที่สนใจในการนำไปใช้ในการจัดการการเปลี่ยนไปใช้ใช้บริการโทรศัพท์เคลื่อนที่กับผู้ให้บริการรายอื่น (Churn Management)

บทที่ 2
การจัดการการยกเลิกบริการโทรศัพท์เคลื่อนที่

การลดลงของจำนวนลูกค้า เป็นปัญหาที่วิกฤติกับธุรกิจในปัจจุบันอย่างมาก ความหมายของการลดลงของจำนวนลูกค้า (Customer turnover) ซึ่งหมายรวมถึง การยกเลิกของลูกค้าในการใช้สินค้าหรือบริการต่อ หรือลูกค้าที่เปลี่ยนจากผู้ให้บริการรายหนึ่งไปยังรายอื่น (Customer Churn) ในความหมายดังกล่าวมักใช้กับบริการโทรคมนาคมเป็นส่วนใหญ่ การที่บริษัทสูญเสียลูกค้าออกไปจากการใช้สินค้าและบริการ แสดงให้เห็นว่าบริษัทต้องแสวงหาลูกค้าจำนวน 4 คนต่อลูกค้าที่ออกไปเพียงคนเดียว หรือหากลูกค้าปัจจุบันยกเลิกการใช้สินค้าหรือบริการปีละ 100 คน บริษัทนั้นก็ต้องแสวงหาลูกค้าใหม่เพื่อชดเชยจำนวน 400 คนต่อเดือน อย่างไรก็ตาม จากการแข่งขันที่รุนแรงแสดงให้เห็นถึง ความไม่คงทนยาวนานของลูกค้าใหม่ กล่าวคือ เป็นการยากที่จะสร้างความน่าสนใจแก่ลูกค้าได้ยืนยาว ยังผลให้การสร้างความสัมพันธ์ระยะยาวและการรักษาลูกค้าเป็นสิ่งสำคัญ (ดร.อนุชิต ศิริกิจ, 2002)

2.1 การยกเลิกของลูกค้าในการใช้บริการ (Customer Churn)

อัตราของลูกค้าที่ยกเลิกการใช้บริการต่อ หรือลูกค้าที่เปลี่ยนจากผู้ให้บริการรายหนึ่งไปยังรายอื่น (Churn Rate) เป็นเรื่องสำคัญสำหรับผู้ให้บริการด้านการสื่อสาร เนื่องจากปัจจุบันในธุรกิจโทรศัพท์เคลื่อนที่มีลูกค้าอยู่น้อยรายที่จะมีความภักดีต่อบริษัท ทั้งนี้เพราะผู้ให้บริการเกือบทั้งหมดมีวิธีการเชิญชวนลูกค้าด้วยคุณลักษณะเด่นที่เพิ่มเติมมาในโทรศัพท์ รวมถึงการแข่งขันกันลดราคาเครื่อง และอัตราค่าบริการ จากผลการสำรวจอัตราเชิร์นในเอเชียโดยเฉลี่ยแล้วจะอยู่ระหว่าง 2-2.5% แต่สามารถ สูงถึง 8%

โดยปัจจัยในการยกเลิกการใช้สินค้า หรือบริการต่อ หรือการเปลี่ยนจากผู้ให้บริการรายหนึ่งไปยังรายอื่นมีมากมาย เช่น

- ความจำเป็น หรือความต้องการในการใช้บริการเปลี่ยนไป
- การย้ายที่อยู่ หรือการเปลี่ยนงาน
- หนี้เสีย หรือไม่ชำระค่าบริการตามใบเรียกเก็บเงิน
- เครือข่ายไม่ครอบคลุม และประสิทธิภาพต่ำ
- อัตราค่าบริการ

- บริการเสริมในรายการส่งเสริมการขายของคู่แข่ง
- ความไม่พอใจในการบริการลูกค้า
- ชื่อเสียงของผู้ให้บริการเสื่อมลง

ปัจจัยดังกล่าวข้างต้นเป็นเพียงส่วนหนึ่งที่มีผลให้ลูกค้ายกเลิกการใช้บริการ หรือเปลี่ยนจากผู้ให้บริการรายหนึ่งไปยังรายอื่น แต่ยังมีปัจจัยอื่นๆ อีกมากมาย ซึ่งปัจจัยที่ทำให้ลูกค้ายกเลิกการใช้บริการของผู้ให้บริการแต่ละรายอาจต่างกัน ขึ้นอยู่กับสภาพแวดล้อมของผู้ให้บริการรายนั้นๆ ด้วย

2.2 การจัดการการยกเลิกบริการโทรศัพท์เคลื่อนที่ (Churn Management in Mobile Service)

ปัจจุบันอัตราการยกเลิกการใช้บริการต่อ หรือลูกค้าที่เปลี่ยนจากผู้ให้บริการรายหนึ่งไปยังรายอื่น (Churn Rate) อยู่ระหว่าง 2-3-เปอร์เซ็นต์ต่อเดือน โดยค่าใช้จ่ายที่ต้องสูญเสียไปเป็นค่าใช้จ่ายในการได้ลูกค้าใหม่สูงกว่าค่าใช้จ่ายในการรักษาลูกค้าเดิมไว้ถึง 5 เท่า สำหรับผู้ให้ที่สามารถลดอัตราการเปลี่ยนไปใช้บริการกับผู้ให้บริการรายอื่นจาก 2 เปอร์เซ็นต์เป็น 1 เปอร์เซ็นต์ จะทำให้ได้รับรายได้ประจำปีเพิ่มขึ้น และเพิ่มส่วนแบ่งทางการตลาดได้มากขึ้น

ด้วยเหตุผลข้างต้นผู้ให้บริการโทรศัพท์เคลื่อนที่จึงตระหนักว่า การเพิ่มรายได้ และลดค่าใช้จ่ายได้ทางหนึ่ง คือการรักษาลูกค้าเดิมให้ใช้บริการต่อไป ดังนั้นผู้ให้บริการหันมาให้ความสนใจกับการจัดการกับการยกเลิกบริการของลูกค้า เพื่อรักษาลูกค้าที่ดีที่สุดไว้

วิธีการจัดการกับการยกเลิกการใช้บริการวิธีหนึ่งคือ เริ่มจากผู้ให้บริการจะต้องเริ่มค้นหาข้อมูลของลูกค้าเก่าที่เคยยกเลิกบริการไปแล้ว และศึกษา และเลือกเครื่องมือที่จะมาช่วยในการหาสาเหตุของการยกเลิกบริการ และนำเครื่องมือนั้นสร้างโมเดลการทำนายจากข้อมูลลูกค้าที่เคยยกเลิกบริการ(Churn) และข้อมูลลูกค้าที่ยังใช้บริการอยู่ (Retain) หลังจากได้โมเดลการทำนายแล้ว นำโมเดลดังกล่าวมาทำนายลูกค้าที่ใช้บริการอยู่ในปัจจุบันว่าเป็นลูกค้าที่มีโอกาสยกเลิกบริการในอนาคตหรือไม่เพื่อวางแผนกลยุทธ์รักษาลูกค้าให้ใช้บริการของบริษัทอย่างต่อเนื่องต่อไป

บทที่ 3
ดาต้าไมนิ่ง

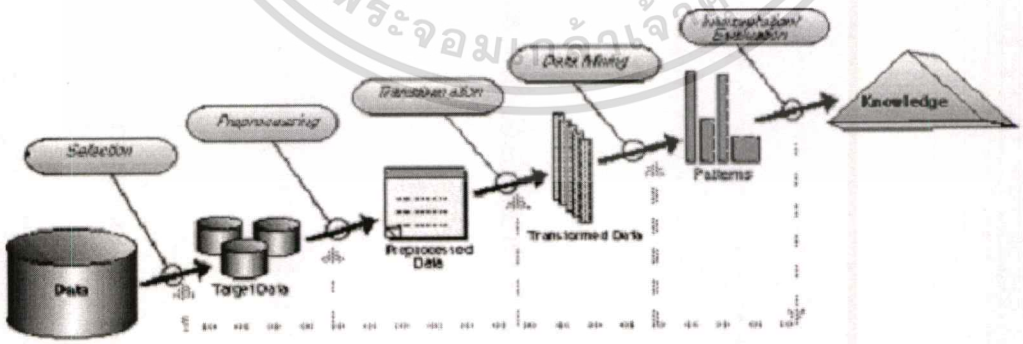
3.1 ดาต้าไมนิ่ง (Data Mining)

ดาต้าไมนิ่ง (Data Mining) หรือการค้นหาความรู้จากฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Database: KDD) คือการนำสารสนเทศที่เป็นประโยชน์และไม่รู้มาก่อนล่วงหน้าจากข้อมูลที่เก็บอยู่ในฐานข้อมูลขนาดใหญ่

ดาต้าไมนิ่งรวมเทคนิคจากเครื่องมือต่างๆ เข้าไว้ด้วยกันเช่นการวิเคราะห์ข้อมูลทางสถิติ (Statistical Data Analysis), การจัดการฐานข้อมูล (Database Management), การเรียนรู้ของเครื่องจักร (Machine Learning) และการแสดงข้อมูลในลักษณะกราฟฟิก (Data Visualization)

3.2 กระบวนการของดาต้าไมนิ่ง (Data Mining Process)

โดยทั่วไปมีการใช้ดาต้าไมนิ่ง และKDDในความหมายเดียวกัน แต่ที่จริงนั้นดาต้าไมนิ่งเป็นเพียงส่วนหนึ่งในกระบวนการใน KDD ดังแสดงในรูปที่ 3.1 และในบทความนี้จะใช้คำว่า ดาต้าไมนิ่ง เพื่อสื่อถึงความหมายของ KDD และดาต้าไมนิ่ง



รูปที่ 3.1 กระบวนการใน Knowledge Discovery from very large Database: KDD (Miguel, 1999)

กระบวนการดาต้าไมนิ่งประกอบด้วยขั้นตอนหลักๆ 5 ขั้นตอนคือ

ขั้นตอนที่ 1 การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)

ขั้นตอนที่ 2 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนที่ 3 การทำดาต้าไมนิ่ง (Data Mining)

ขั้นตอนที่ 4 การวิเคราะห์ผล (Analysis of Results)

ขั้นตอนที่ 5 การนำความรู้มาใช้ (Assimilation of Knowledge)

โดยวัตถุประสงค์ทางธุรกิจนั้นจะใช้จนตลอดทุกขั้นตอนของดาต้าไมนิ่ง ถึงแม้ว่าแต่ละขั้นตอนจะมีการทำตามลำดับข้างต้น แต่ก็อาจมีการย้อนกลับมาทำในแต่ละขั้นตอนใหม่ อีกครั้ง เวลาที่ใช้ในแต่ละขั้นตอนนั้นไม่เท่ากัน ขั้นตอนในการเตรียมข้อมูลจะใช้เวลาถึง 60 เปอร์เซ็นต์ของเวลาในการทำดาต้าไมนิ่งทั้งหมด แต่ขั้นตอนในการทำดาต้าไมนิ่งนั้นใช้เวลาเพียง 10 เปอร์เซ็นต์เท่านั้นโดยแต่ละขั้นตอนที่รายละเอียดต่างๆ ดังต่อไปนี้

3.2.1 การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)

เป็นขั้นตอนสำคัญในกระบวนการของดาต้าไมนิ่ง การกำหนดวัตถุประสงค์ให้ชัดเจนต้องอาศัยความเข้าใจในปัญหาหรือโอกาสทางธุรกิจ หากกำหนดวัตถุประสงค์ไม่ชัดเจน อาจทำให้ผลที่ได้มีความคลุมเครือ ไม่สามารถนำไปใช้ได้ ต้องกลับไปเริ่มต้นในขั้นตอนแรกใหม่อีกครั้ง ตัวอย่างการกำหนดวัตถุประสงค์ที่ชัดเจนเช่น การกำหนดวัตถุประสงค์ในการพัฒนากลยุทธ์ทางการตลาดเพื่อรักษาส่วนแบ่งทางการตลาดเดิมของโทรศัพท์เคลื่อนที่ในแบบชำระค่าบริการก่อน (Prepaid) ในเขตกรุงเทพมหานครระหว่างเดือนมกราคมถึงเมษายน เป็นต้น

3.2.2 การเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลามากกว่าขั้นตอนอื่นๆ ซึ่งประกอบด้วยขั้นตอนย่อย 3 ขั้นตอนดังนี้

1) การเลือกข้อมูล (Data Selection)

เป็นการกำหนดแหล่งข้อมูลต่างๆ ทั้งในและนอกองค์กร และเลือกข้อมูลที่จำเป็นเพื่อนำไปใช้ในการวิเคราะห์เบื้องต้นในขั้นตอนการเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่งในขั้นตอนถัดไป ข้อมูลที่เลือกจะผันแปรตามวัตถุประสงค์ทางธุรกิจ โดยตัวแปรที่เลือกมาจะมีชนิด(type), ค่า(value), รูปแบบ(format), และลักษณะ(characteristic)ที่ชัดเจน ซึ่งชนิดของข้อมูลแบ่งได้เป็น 2 ประเภทคือ

- ข้อมูลที่เป็นจำนวนหรือตัวเลข (Quantitative data) โดยแบ่งย่อยได้เป็นแบบจำนวนเต็ม (discrete) เช่นจำนวนพนักงาน และแบบจำนวนจริง (continuous) เช่น รายได้, ค่าเฉลี่ยต่างๆเป็นต้น
- ข้อมูลที่แบ่งเป็นกลุ่ม (Categorical data) โดยแบ่งย่อยออกได้เป็นแบบมีลำดับ (Ordinal categorical data) เช่นระดับเครดิตลูกค้า (ดี, ปานกลาง, ไม่ดี) และแบบไม่มีลำดับ (Nominal categorical data) เช่น สถานภาพสมรส (โสด, แต่งงาน, หย่า) เป็นต้น

2) การเตรียมข้อมูลก่อนประมวลผล (Data Preprocessing)

เป้าหมายในการเตรียมข้อมูลก่อนนำไปประมวลผล คือได้ข้อมูลที่มีคุณภาพ, ถูกต้อง และเป็นปัจจุบัน โดยเริ่มจากการพิจารณาปริมาณ โครงสร้าง และความเป็นไปได้ของข้อมูลที่จะทำให้เกิดการทำดาต้าไมนิ่งได้ เนื่องจากข้อมูลที่เลือกถูกรวบรวมมาจากหลายแหล่ง หรือข้อมูลไม่สอดคล้องกันเป็นต้น และหากข้อมูลที่เลือกไม่เป็นปัจจุบันจะมีผลทำให้เกิดข้อมูลที่มีความถูกต้องและความคุณภาพต่ำตลอดทั้งกระบวนการในดาต้าไมนิ่ง ซึ่งในระหว่างขั้นตอนนี้สิ่งที่เกิดขึ้นบ่อยๆ คือ

- ข้อมูลมีความคลาดเคลื่อน(Noisy data)

  คือตัวแปรมีค่าคลาดเคลื่อนไป และค่าคลาดเคลื่อนที่เกิดขึ้นเรียกว่า Outlier ซึ่งอาจเป็นข้อมูลที่มีประโยชน์ทำให้เห็นโอกาสใหม่ๆ หรือเป็นข้อมูลที่ไม่มีประโยชน์ไม่สามารถนำไปใช้ได้ เช่น การเก็บข้อมูลอายุ 300 ปี (ซึ่งในความเป็นจริงอายุเฉลี่ยโดยทั่วไปอยู่ในช่วงประมาณ 70 ถึง 90 ปี) หรือรายได้ที่มีค่าติดลบ (ซึ่งเป็นค่าที่เป็นไปไม่ได้) เป็นต้น โดยต้องทำการแก้ไขข้อมูลที่ผิดพลาดเหล่านั้นให้ถูกต้อง หรือตัดค่านั้นทิ้งไปด้วยวิธีต่างๆเช่น ทำการจัดเป็นกลุ่ม (Clustering) หรือการทำ Regression หรือวิธีการ Binning โดยทำการเรียงลำดับข้อมูลและแบ่งข้อมูลเป็นช่วงเท่ากันแล้วนำค่ากลางมาใช้ เป็นต้น เพื่อให้ข้อมูลมีความถูกต้องสมเหตุสมผล

- ค่าของข้อมูลขาดหายไป (Missing value)

  คือค่าไม่ปรากฏในข้อมูลที่เลือกมา และค่าคลาดเคลื่อนที่ใช้ไม่ได้ซึ่งเราตัดทิ้งไป มีวิธีแก้ไขได้หลายวิธี เช่นหากข้อมูลที่ขาดหายไปมีจำนวนน้อยสามารถแก้ไขได้โดยตัดข้อมูลเหล่านั้นทิ้งไป หรือการเติมข้อมูลที่ขาดหายไปด้วยการหาค่าเฉลี่ยหรือการทำนายค่าเป็นต้น

3) การแปลงข้อมูล (Data Transformation)

ข้อมูลบางชนิดเป็นข้อมูลที่ไม่เหมาะสมกับการนำมาวิเคราะห์ จึงต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปใช้ในดาต้าไมนิ่งอัลกอริทึมเทคนิคการแปลงข้อมูลมีวิธีทำทั้งในการแปลงรูปแบบของข้อมูลแบบง่ายๆ ไปจนถึงการใช้เครื่องมือในการลดจำนวนข้อมูลทางสถิติที่ซับซ้อนด้วยการรวมตัวแปรหลายตัวเป็นตัวแปรตัวเดียว เพื่อลดจำนวนตัวแปรสำหรับการประมวล เช่นการประเมินความสนใจของกลุ่มลูกค้าเป้าหมายในสินค้าตัวใหม่ สามารถที่จะรวมตัวแปรหลายตัวที่เกี่ยวข้องกันอาทิเช่น รายได้, รหัสไปรษณีย์ และระดับการศึกษา ได้เป็นตัวแปรหนึ่งตัวที่แสดงถึงความสนใจของกลุ่มลูกค้าเป้าหมาย เป็นต้น และตัวอย่างในการแปลงข้อมูลแบบง่ายๆ เช่น ชนิดของโทรศัพท์เคลื่อนที่ Nokia, Ericson, Siemens, Motorola ต้องแปลงให้อยู่ในรูปแบบอื่น โดยอาจกำหนดให้ Nokia = 001, Ericson = 010, Siemens = 100, Motorola = 101 เป็นต้น

3.2.3 การทำดาต้าไมนิ่ง (Data Mining)

ขั้นตอนนี้เป็นขั้นตอนที่เป็นดาต้าไมนิ่งที่แท้จริง โดยในการเลือกใช้อัลกอริทึมนั้นจะขึ้นอยู่กับวัตถุประสงค์ที่วางไว้ในขั้นตอนแรก และนำข้อมูลที่เตรียมไว้มาใช้ในอัลกอริทึมที่เลือก สิ่งต่างๆ ที่เกิดขึ้นระหว่างขั้นตอนการทำดาต้าไมนิ่งจะแปรผันตามแอปพลิเคชั่นที่ทำการพัฒนา ตัวอย่างเช่น ในกรณีของ การแบ่งกลุ่มฐานข้อมูล (Database Segmentation) นั้น การรันอัลกอริทึมเพียงหนึ่ง หรือสองรอบ ก็จะได้ผลลัพธ์ และดำเนินไปยังขั้นตอนการวิเคราะห์เป็นลำดับต่อไป แต่ถ้าเป็นโมเดลในการทำนาย (Predictive Modeling) นั้นจะทำการวนลูปซ้ำในการสร้างโมเดลหลายรอบ ถึงจะสามารถนำมาใช้กับข้อมูลจริงได้ กล่าวคือแต่ละอัลกอริทึมนั้นมีข้อดี และข้อเสียที่แตกต่างกันไป การเลือกอัลกอริทึมที่เหมาะสมจึงขึ้นอยู่กับปัจจัยหลายประการ เช่น วัตถุประสงค์ของธุรกิจ, ความสามารถในการรองรับชนิดของข้อมูลของอัลกอริทึมนั้นๆ, ความสามารถในการอธิบายเอาท์พุต, ความสามารถในการปรับเปลี่ยนขนาดของการรองรับข้อมูลได้ (scalability), ง่ายต่อการนำมาประยุกต์ใช้งาน

Predictive modeling, Database segmentation, Link analysis และ deviation detection เป็น 4 โอเปอร์เรชั่นหลักๆ ที่ใช้ในการนำมาพัฒนาแอปพลิเคชั่น มีรายละเอียดดังต่อไปนี้

1) Predictive Modeling

Prediction modeling มีคุณสมบัติบางอย่างเหมือนกับประสบการณ์การเรียนรู้ของมนุษย์ โดยใช้การสังเกตสิ่งๆ หนึ่ง และนำมาสร้างรูปแบบโมเดลตามลักษณะที่สำคัญของสิ่งนั้นๆ ในดาต้าไมนิ่งนั้นใช้ predictive model ในการวิเคราะห์ฐานข้อมูลที่มีอยู่เพื่อบ่งชี้ลักษณะที่จำเป็นเกี่ยวกับข้อมูล

นั้น ซึ่งข้อมูลจะต้องมีความสมบูรณ์ โดยจะสังเกตได้จากความสามารถในการทำนายได้อย่างถูกต้องของโมเดล โมเดลจะต้องให้คำตอบที่ถูกต้องตรงกับคำตอบของสิ่งที่ได้พิสูจน์แล้ว ก่อนที่จะเริ่มนำมาใช้ในการทำนายจริง ซึ่งวิธีการแบบนี้เรียกว่า Supervised Learning

Prediction Model เป็นโมเดลในการวิเคราะห์ข้อมูลที่มีอยู่เพื่อทำนายแนวโน้มของข้อมูลที่จะเกิดขึ้นในอนาคต โดยในการสร้างโมเดลประกอบด้วย 2 ช่วงคือ Training และ Testing ซึ่ง Training เป็นช่วงของการสร้างโมเดลใหม่โดยใช้ข้อมูลเก่าที่มีอยู่แล้ว และ Testing คือการใช้ข้อมูลที่ไม่เคยใช้ในการสร้างโมเดลมาทำการทดสอบความถูกต้องของโมเดล และโมเดลสามารถเป็นได้ทั้งกลุ่มคำสั่งของ SQL, เงื่อนไข IF THEN, หรือกลุ่มคำสั่งภาษาซี และPrediction Modeling แบ่งออกเป็น 2 แบบดังนี้คือ

- Classification

  การแบ่งข้อมูลออกเป็นกลุ่ม และใช้ Predictive Model ทำนายว่าข้อมูลควรอยู่ในกลุ่มใด เช่นในการให้เงินกู้ จะทำนายว่าลูกค้าควรอยู่ในกลุ่มลูกค้าชั้นใด

- Value prediction (Forecasting)

  ใช้ Predictive Model ในการทำนายค่าที่สัมพันธ์กับข้อมูลที่มีอยู่ เช่นการพยากรณ์อากาศ หรือการทำนายหุ้นเป็นต้น

2) Database Segmentation (Clustering)

เป็นวิธีการนำข้อมูลมาแบ่งเป็นกลุ่มที่มีความสัมพันธ์กัน เพื่อวิเคราะห์หาลักษณะที่เหมือน หรือความแตกต่างของข้อมูลในแต่ละกลุ่ม อัลกอริทึมการแบ่งกลุ่ม (Segmentation Algorithm) สามารถทำการแบ่งกลุ่มได้เองโดยไม่ต้องอาศัยคนมาทำการกำหนดว่าจะแบ่งข้อมูลออกเป็นกี่กลุ่ม หรือแบ่งกลุ่มของข้อมูลในลักษณะใด

การแบ่งกลุ่มของข้อมูล (Segmentation) สนับสนุนแอปพลิเคชั่นทางธุรกิจต่างๆเช่น ข้อมูลประวัติของลูกค้า (customer profile) หรือกลุ่มตลาดเป้าหมาย (target marketing) และการรักษาลูกค้าให้ใช้บริการต่อไป (customer retention)

3) Link Analysis

เป็นการค้นหาความสัมพันธ์ (Associations) ระหว่างข้อมูล หรือกลุ่มข้อมูล เช่นการหาความสัมพันธ์ระหว่างสินค้า หรือบริการที่ลูกค้าชอบซื้อร่วมกัน, การซื้อสินค้าประเภทหนึ่ง แล้วจะซื้อสินค้าอีกประเภทหนึ่งต่อเนื่องกัน เป็นต้น แบ่งเป็น 3 แบบคือ

- Association Discovery

ใช้ในการวิเคราะห์การซื้อสินค้าเพื่อหาความสัมพันธ์ที่ซ่อนอยู่ระหว่างผลิตภัณฑ์ซึ่งจะขายได้ดีเมื่อขายคู่กัน การวิเคราะห์ในลักษณะนี้ถูกเรียกว่า Market Basket Analysis

- Sequential Pattern Discovery

ใช้ในการกำหนดความสัมพันธ์ในการซื้อสินค้าที่ไม่มีความเกี่ยวข้องกันในช่วงเวลาหนึ่งๆ ที่มีการแสดงข้อมูลเป็นลำดับในการซื้อสินค้าและใช้บริการ เช่นเมื่อซื้อพัดลมแล้วต่อมาอาจจะซื้อแอร์มาใช้ เป็นต้น ช่วยทำให้เข้าใจพฤติกรรมการซื้อของลูกค้า และนำมาจัดรายการส่งเสริมการขาย

- Similar Time Sequence Discovery

เป็นการค้นหาความสัมพันธ์ระหว่างข้อมูลสองกลุ่มในช่วงระยะเวลาหนึ่งเช่นรายเดือน ,รายปี โดยเทียบระดับความเหมือนกันระหว่างแบบสองแบบ(Patterns) ในช่วงระยะเวลาที่ทำการทดลองเดียวกัน

4) Deviation Detection

เป็นการวิเคราะห์ว่ามีอะไรแตกต่างจากกลุ่มอื่น โดยใช้กราฟ หรือรูปภาพ เพื่อแสดงให้เห็นความแตกต่างจากกลุ่มอื่น แอปพลิเคชั่นที่อัลกอริทึมนี้สนับสนุนมีทั้ง การป้องการโกง (Fraud detection) ในการใช้บัตรเครดิต ในสินไหมการประกัน หรือในการใช้บัตรโทรศัพท์ และการควบคุมคุณภาพ

3.2.4 การวิเคราะห์ผล (Analysis of Results)

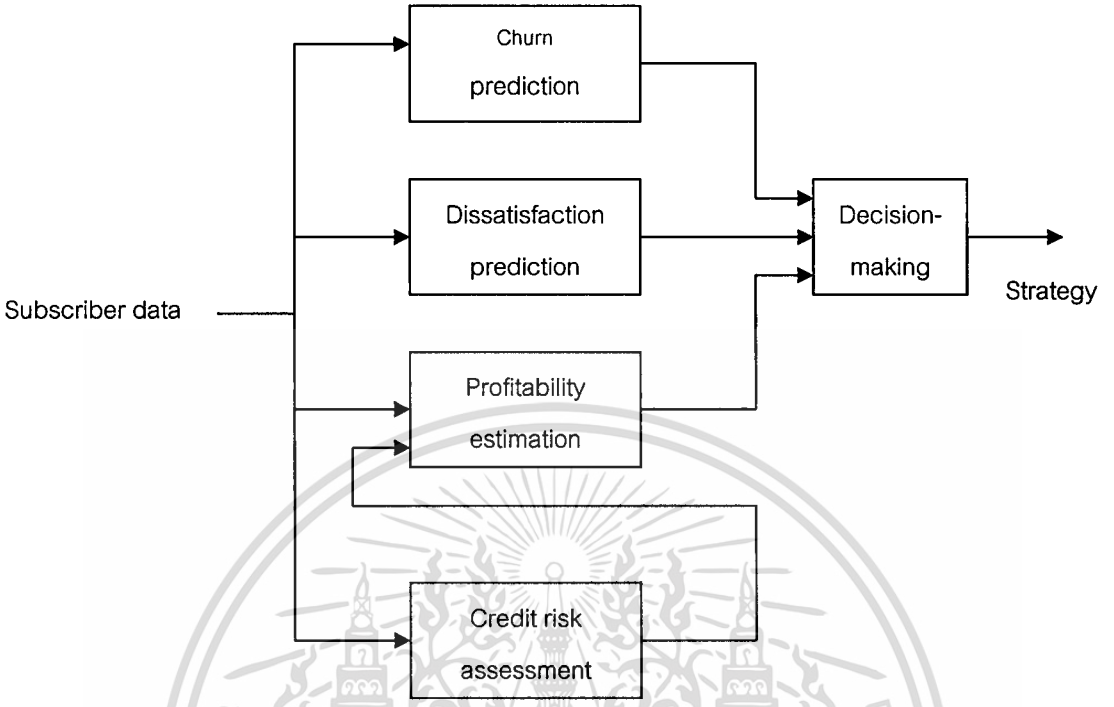
นำผลที่ได้จากการทำดาต้าไมนิ่งมาแปล และตีความหมาย ถ้าผลที่ได้ไม่เป็นไปตามวัตถุประสงค์ที่วางไว้ โดยพิจารณาว่าเป็นเพราะขั้นตอนใด เพื่อย้อนกลับไปแก้ไขในขั้นตอนนั้นๆ โดยการวิเคราะห์ผลในขั้นตอนนี้ขึ้นอยู่กับชนิดของแอปพลิเคชั่นที่ทำการพัฒนา

3.2.5 การนำความรู้มาใช้ (Assimilation of Knowledge)

นำสารสนเทศที่ได้ไปใช้ตามวัตถุประสงค์ที่วางไว้ ซึ่งสารสนเทศใหม่ที่ได้นั้นต้องถูกต้อง (Valid) และนำมาใช้ได้ (Actionable) โดยสารสนเทศที่ได้จะช่วยให้เห็นแนวทางการทำธุรกิจในรูปแบบใหม่ๆ หรือช่วยในเกิดประโยชน์สูงสุดเพื่อนำไปพัฒนาองค์กร

3.3 ดาต้าไมนิ่ง กับการจัดการการยกเลิกบริการ (Data Mining and Churn Management)

เทคนิคในดาต้าไมนิ่งบางเทคนิคสามารถนำมาใช้ในการทำนายสาเหตุ และกลุ่มลูกค้าที่มีโอกาสยกเลิกบริการ หรือเปลี่ยนจากผู้ให้บริการรายหนึ่งไปยังผู้ให้บริการรายอื่น เทคนิคต่างๆ ได้แก่ logic regression, decision trees, neural networks และ boosting โดยเทคนิคหนึ่งที่นิยมนำมาใช้คือดิซิชั่นทรี (Decision Tree) เนื่องจากดิซิชั่นทรีนั้นไม่เพียงแต่สามารถบอกค่าในการทำนายการเปลี่ยนไปบริการระบบอื่นของแต่ละเบอร์โทรศัพท์ แต่ยังบอกถึงสาเหตุของการยกเลิกการใช้บริการกับระบบ



รูปที่ 3.2 โครงร่างสำหรับการทำนายการยกเลิกบริการ และการเพิ่มผลกำไรสูงสุด (Michael, 2000)

ซึ่งดิซิชั่นทรีเป็น Predictive Modeling ที่เป็นโอเปอร์เรชั่นหนึ่งในดาต้าไมนิ่งที่ใช้ในการแบ่งกลุ่ม และการทำนาย โดยใช้ข้อมูลของลูกค้าที่ใช้บริการโทรศัพท์เคลื่อนที่ที่มีอยู่ในองค์กรมาทำนายกลุ่มลูกค้า และสาเหตุการเปลี่ยนไปใช้บริการกับผู้ให้บริการรายอื่น นอกจากนี้ยังสามารถนำข้อมูลดังกล่าวมาประมาณผลกำไรที่จะได้รับในแต่ละเดือน และการประมินความเสี่ยง ความเชื่อถือหรือความไว้วางใจของลูกค้า (ซึ่งมีอิทธิพลต่อผลประโยชน์ที่จะได้รับในแต่ละเดือน) ได้อีกด้วย และเมื่อนำข้อมูลมาใช้ในการสร้างโมเดลแล้ว โมเดลการทำนายที่ได้สามารถนำมาใช้ในการสนับสนุนการตัดสินใจ เพื่อเป็นแนวทางในการวางกลยุทธ์ทางการตลาด ดังแสดงในรูปที่ 3.2

3.4 กระบวนการดาต้าไมนิ่ง กับการจัดการการยกเลิกบริการ (Data Mining Process and Churn Management)

ดาต้าไมนิ่งเป็นอีกวิธีหนึ่งที่นิยมนำมาใช้ร่วมกับการจัดการการยกเลิกบริการ หรือการเปลี่ยนไปใช้บริการกับผู้ให้บริการระบบอื่น และในบทความนี้ได้นำเสนอการนำดิซิชั่นทรีมาใช้ในการสร้างโมเดลในการทำนาย และก่อนการสร้างโมเดลดังกล่าวนั้นจะต้องผ่านกระบวนการดาต้าไมนิ่งดังต่อไปนี้

3.4.1 วัตถุประสงค์ทางธุรกิจ (Business Objective)

วัตถุประสงค์การทำดาต้าไมนิ่งในธุรกิจการบริการโทรศัพท์เคลื่อนที่ ในการจัดการกับการเปลี่ยนไปใช้บริการกับผู้ให้บริการระบบอื่น เพื่อวิเคราะห์หา

- สาเหตุของการเปลี่ยนไปใช้บริการกับระบบอื่น
- กลุ่มลูกค้าที่มีโอกาสยกเลิกการใช้บริการ หรือเปลี่ยนไปใช้บริการกับผู้ให้บริการรายอื่น

3.4.2 เตรียมข้อมูล (Prepare Data)

ในขั้นตอนการเตรียมข้อมูลนี้เป็นขั้นตอนหนึ่งที่มีความสำคัญ และใช้เวลานานเนื่องจากจะต้องพิจารณาเลือกข้อมูลที่จะนำมาใช้ในการวิเคราะห์ให้เหมาะสมกับวัตถุประสงค์ที่ตั้งไว้ในตอนแรก ซึ่งประกอบด้วยขั้นตอนย่อยๆ ได้แก่ ขั้นตอนการเลือกข้อมูล, การเตรียมข้อมูลก่อนประมวลผล และการแปลงข้อมูล

ในธุรกิจการให้บริการระบบโทรศัพท์เคลื่อนที่ ข้อมูลที่นำมาใช้ทำดาต้าไมนิ่ง เป็นข้อมูลของ Subscriber โดยสามารถแบ่งเป็นประเภทหลักๆ ได้ดังต่อไปนี้

- กลุ่มเครือข่าย (Network) : เรคคอร์ดรายละเอียดการโทร (วัน, เวลา, และพื้นที่โทรออก เป็นต้น)
- กลุ่มใบเรียกเก็บเงิน (Billing) : ข้อมูลทางการเงินในใบเรียกเก็บเงิน (ค่าบริการรายเดือน, วงเงินใช้บริการ, ยอดค่าใช้บริการ เป็นต้น)
- กลุ่มข้อมูลการรับบริการ (Application for Service) : รายละเอียดในการติดต่อระหว่างลูกค้ากับผู้ให้บริการระบบ, รายงานเกี่ยวกับเครดิตของลูกค้า
- กลุ่มข้อมูลการตลาด (Market) : รายละเอียดของอัตราค่าบริการ, โฆษณาและโปรโมชั่นของผู้ให้บริการรายอื่นๆ
- กลุ่มข้อมูลลูกค้า (Customer) : ระบบโทรศัพท์ที่ใช้, พื้นที่ที่จดทะเบียน, จำนวนปีที่ใช้บริการ, วันที่จดทะเบียน เป็นต้น
- กลุ่มข้อมูลประชากร (Demographic) : อายุ, อาชีพ, เพศ, สถานภาพการสมรสม, รายได้ เป็นต้น

ข้อมูลที่ใช้เป็นข้อมูลลูกค้า โดยพิจารณาข้อมูลเป็นรายเบอร์โทรศัพท์ (Subscriber: หนึ่งเบอร์โทรศัพท์ นับเป็นหนึ่ง Subscriber) เช่น ประเภทของระบบโทรศัพท์, พื้นที่ที่จดทะเบียน, จำนวนปี ณ ปัจจุบันที่ใช้บริการกับระบบ,วงเงินสูงสุดที่ได้รับ เป็นต้น และใช้ข้อมูลเกี่ยวกับการใช้โทรศัพท์ของลูกค้า เช่น ชนิดของโปรโมชั่นที่ลูกค้าใช้, จำนวนครั้งที่ลูกค้าจ่ายค่าบริการล่าช้า, ค่าใช้โทรศัพท์

เฉลี่ยสามเดือนสุดท้าย เป็นต้น โดยนำข้อมูลของลูกค้าที่เปลี่ยนไปใช้บริการกับระบบอื่นไปแล้วกับข้อมูลของลูกค้าปัจจุบันที่ยังคงใช้บริการกับระบบอยู่มาสร้างโมเดลในการทำนาย และนำข้อมูลของลูกค้าปัจจุบันที่ยังคงใช้บริการกับระบบอยู่มาทำนายว่าลูกค้ากลุ่มใดที่มีโอกาสยกเลิกบริการหรือเปลี่ยนไปใช้บริการกับผู้ให้บริการรายอื่น ตัวอย่างแอตทริบิวต์ที่เลือกมาใช้ในการทำดาต้าไมนิ่งเช่น

- Customer Age : อายุของลูกค้า
- Subscriber Age : อายุการใช้งาน
- Area of Subscriber : พื้นที่ที่จดทะเบียน
- Credit Limit : วงเงินใช้บริการ
- Promotion Code : รหัสโปรโมชั่น
- Number of Not Paid : จำนวนครั้งที่เคยไม่ชำระค่าบริการ
- Number of Over Credit Limit : จำนวนครั้งที่ใช้เกินวงเงินบริการ
- Last 3 month Average Invoice : ค่าบริการเฉลี่ย 3 เดือนสุดท้าย

และหลังจากเลือกข้อมูลได้แล้วจะต้องทำการเตรียมข้อมูลก่อนนำไปประมวลผล ซึ่งข้อมูลที่ได้เลือกมานั้นเป็นข้อมูลที่ถูกต้อง และเป็นปัจจุบัน เนื่องจากข้อมูลเป็นข้อมูลที่มาจากแหล่งข้อมูลภายในองค์กรเพียงแหล่งเดียว และยังเป็นข้อมูลที่นำมาจากคลังข้อมูล (Data Warehouse) ขององค์กรดังนั้นจึงไม่ต้องทำการเตรียมข้อมูลก่อนนำไปประมวลผล และในทำการแปลงข้อมูลให้เหมาะสมในการนำมาวิเคราะห์นั้นข้อมูลที่เลือกมาใช้มีค่าที่มีความเหมาะสมที่จะนำไปใช้วิเคราะห์ได้เลย ดังนั้นจึงสามารถนำข้อมูลที่เลือกมาใช้ในขั้นตอนการทำดาต้าไมนิ่งต่อไป

บทที่ 4

SLIQ Classification

SLIQ (Supervised Learning In Quest) เป็นการแบ่งกลุ่มแบบดิซิชั่นทรีที่สามารถรองรับแอตทริบิวต์ที่เป็นทั้งแบบตัวเลข (Numerical) และแบบกลุ่ม (Categorical) และยังขยายความสามารถในการรองรับกลุ่มของข้อมูลที่มีขนาดใหญ่มากได้ด้วย โดยในขั้นตอนการสร้างต้นไม้ (Growing Phase) จะมีการใช้เทคนิค Pre-sorting ซึ่งช่วยลดจำนวนครั้งในการหาจุดแบ่งที่ดีที่สุด (Best Split) บนแอตทริบิวต์แบบตัวเลขและยังสามารถทำการเรียงข้อมูลบนดิสก์หากข้อมูลมีขนาดใหญ่เกินกว่าที่จะทำการรันบนหน่วยความจำได้ ดังนั้นจึงเลือก SLIQ มาเป็นอัลกอริทึมในการพัฒนาระบบ

4.1 ซุปเปอร์ไวส์เลินนิ่ง (Supervised Learning)

Supervised Learning คือการเรียนรู้ด้วยตัวอย่าง (example) โดยใช้ตัวอย่างหาโครงร่าง (Pattern) ที่แยกความแตกต่างระหว่าง ค่าในคอลัมน์ที่ขึ้นกับคอลัมน์อื่น (dependent column) โดย classification, estimation และ prediction เป็นแบบซุปเปอร์ไวส์เลินนิ่ง

เริ่มจากสร้างกลุ่มข้อมูล (Data set) ที่เป็นตัวอย่าง (example) ที่ประกอบด้วยหลายๆ คอลัมน์ที่เป็นข้อมูลเกี่ยวกับลูกค้า เรียกว่า คอลัมน์ที่ไม่ขึ้นกับคอลัมน์อื่น (independent column) และเพิ่มคอลัมน์ที่บอกว่าลูกค้าจะยกเลิกบริการ (Churn) หรือ ยังคงใช้บริการต่อไป (Retain) เรียกว่า คอลัมน์ที่ขึ้นกับคอลัมน์อื่น (dependent column) หรือคอลัมน์ที่เราสนใจ โดยคอลัมน์ที่ขึ้นอยู่กับคอลัมน์อื่นเป็นคอลัมน์ที่เก็บลาเบล (Label) เช่น Churn หรือ Retain ที่จะทำนายจากคอลัมน์ที่ไม่ขึ้นกับคอลัมน์อื่น และหลังจากได้โมเดลแล้ว ทำให้สามารถทำนายได้ว่าลูกค้าคนนั้นๆ มีโอกาสจะยกเลิกบริการ (Churn) หรือไม่ โดยที่เราไม่รู้ค่าในคอลัมน์ที่ขึ้นกับคอลัมน์อื่นมาก่อน

ซุปเปอร์ไวส์เลินนิ่งสามารถเรียนรู้ได้เมื่อข้อมูลตัวอย่างนั้นมีลาเบลที่แบ่งกลุ่มไว้ก่อนแล้ว แต่ถ้าไม่มีก็จะทำให้ไม่สามารถเรียนรู้ได้ ซึ่งต่างจากอันซุปเปอร์ไวส์เลินนิ่ง (Unsupervised learning) ซึ่งไม่ได้ใช้ประโยชน์จากการเรียนรู้ด้วยข้อมูลตัวอย่าง กลับกันอันซุปเปอร์ไวส์เลินนิ่งพยายามที่จะหาโครงร่าง (Pattern) ซึ่งอยู่ภายในข้อมูล แต่ไม่มีลาเบลในกลุ่มข้อมูลตัวอย่าง และสามารถทำการแบ่งกลุ่มข้อมูลตัวอย่าง จนกระทั่งได้กลุ่มที่มีสมาชิกที่เหมือนกันอยู่ในกลุ่มเดียวกัน อัลกอริทึมที่เป็นอันซุปเปอร์ไวส์เลินนิ่ง เช่น Neural network เป็นต้น

4.2 การแบ่งกลุ่มแบบดิซิชั่นทรี (Decision Tree Classification)

การแบ่งกลุ่ม (Classification) คือขบวนการของการใส่แต่ละแถว (row) ลงในกลุ่มที่แบ่งไว้แล้ว ตัวอย่างของการแบ่งกลุ่ม (Classification) เช่น การแบ่งกลุ่มลูกค้าเป็น ลูกค้าที่มีโอกาสยกเลิกบริการ (Churn) หรือยังคงใช้บริการต่อไป (Retain) เราสามารถใช้การแบ่งกลุ่ม (Classification) ในการสร้างโมเดลเพื่อแบ่งกลุ่มลูกค้าที่ใช้บริการอยู่ในปัจจุบัน และใช้โมเดลทำนายว่าลูกค้ามีโอกาสยกเลิกบริการในอนาคตหรือไม่

ดิซิชั่นทรี (Decision tree) เป็นไดอะแกรม หรือโฟรชาร์ทซึ่งแสดงการแบ่งกลุ่ม (Classification) หรือโมเดลในการทำนาย (Predictive Model) และเป็นโมเดลที่ใช้สำหรับทำนายคอลัมน์ที่ขึ้นกับคอลัมน์อื่นจำนวนหนึ่งคอลัมน์(Single Dependent Column) โครงสร้างของ ดิซิชั่นทรีเป็นโครงสร้างที่เป็นลำดับของคำถามง่ายๆ และคำตอบจะได้จากการท่องไปตามเส้นทางในแนวดิ่งลงของต้นไม้ เมื่อท่องไปจนถึงจุดปลายสุด หรือใบของต้นไม้ (leaf node) ก็จะได้ค่าของคอลัมน์ที่ขึ้นกับคอลัมน์อื่นซึ่งอยู่ในลีฟโหนดนั้นๆ

เหตุผลในการเลือกใช้ดิซิชั่นทรีในการพัฒนาระบบนี้ เพราะดิซิชั่นทรีนั้นให้ผลการทำนายถึงสาเหตุที่ลูกค้ายกเลิกการใช้บริการโทรศัพท์เคลื่อนที่ นอกจากเหตุผลหลักนี้ ดิซิชั่นทรียังมีข้อดีอีกหลายอย่างเช่น

- สามารถสร้างได้รวดเร็วกว่าเมื่อเทียบกับวิธีอื่น
- เป็นโมเดลที่สามารถเข้าใจได้ง่าย
- สามารถแปลงไปเป็น SQL Statement ได้ง่าย
- ให้ผลที่ถูกต้องเมื่อเปรียบเทียบกับวิธีการ Classification แบบอื่น

ขบวนการในการสร้างโมเดลของ Decision tree classifier ส่วนใหญ่จะแบ่งเป็น 2 เฟส คือ การสร้างต้นไม้ (Tree Building) และ การตัดกิ่ง (Tree Pruning)

4.2.1 การสร้างต้นไม้ (Tree Building)

ในการสร้างต้นไม้นั้น ข้อมูลจะนำมาทำการแบ่งกลุ่มแบบวนซ้ำไปเรื่อย ๆ จนกระทั่งสมาชิกทุกตัว หรือสมาชิกเกือบทุกตัว (จำนวนของสมาชิกสูงสุดขึ้นอยู่กับการกำหนดของผู้สร้างต้นไม้) ในแต่ละกลุ่มจะอยู่ในกลุ่ม (Class) เดียวกัน โดยใช้ข้อมูลเดิมที่มีอยู่ซึ่งเป็นส่วนสำหรับสร้างต้นไม้ (Training set) มาทำการแบ่งเป็นสองกลุ่มหรือมากกว่าโดยใช้แอตทริบิวต์ที่มีค่าที่ชี้การแบ่งกลุ่ม (Splitting index) มากที่สุดมาใช้แบ่งกลุ่ม ซึ่งรูปแบบที่ใช้ในการแบ่งกลุ่มข้อมูลขึ้นอยู่กับชนิดของแอตทริบิวต์ โดยวิธีการคำนวณค่าที่ชี้การแบ่งกลุ่ม (Splitting index) และรูปแบบของแอตทริบิวต์ที่ใช้ในการแบ่งกลุ่มได้อธิบายไว้ในหัวข้อ 4.3.2 และในรูปที่ 4.1 แสดงอัลกอริทึมในการสร้างต้นไม้

(Tree-Building Algorithm) โดย T คือข้อมูลที่จะนำมาใช้ในการสร้างต้นไม้, S คือกลุ่มข้อมูลที่เกิดจากการแบ่งข้อมูล T ออกเป็นสองส่วน และ A คือแอตทริบิวต์ที่นำมาคำนวณค่าที่ชี้การแบ่งกลุ่ม (splitting index)

```
MakeTree(Training Data T)
        Partition (T);

Partition(Data S)
        If (all points in S are in the same class)) then return;
        Evaluate splits for each attribute A
        Use best split found to partition S into S1 and S2;
        Partition(S1);
        Partition(S2);
```

รูปที่ 4.1 อัลกอริทึมในการสร้างต้นไม้ (Tree-Building Algorithm)

โดยในการสร้างต้นไม้มีขั้นตอนหลัก 2 ขั้นตอนในระหว่างสร้างต้นไม้คือ

1) การประเมินค่าที่ชี้การแบ่งกลุ่ม (Splitting index) สำหรับแต่ละแอตทริบิวต์และเลือกค่าของการแบ่งกลุ่มที่ดีที่สุด (best split)
2) ทำการแบ่งกลุ่มโดยใช้ค่าของการแบ่งกลุ่มที่ดีที่สุด (Best split)

4.2.2 การตัดกิ่ง (Tree Pruning)

หลังจากสร้างต้นไม้เสร็จจนได้ต้นไม้ที่สมบูรณ์แล้ว ขั้นตอนต่อไปเป็นการตัดกิ่ง (Prune) ของต้นไม้ คือเลือกตัดกิ่งที่มีข้อมูลที่ผิดพลาด และมีค่าแกว่งไปแกว่งมา ซึ่งกิ่งเหล่านี้สามารถนำไปสู่ความผิดพลาดเมื่อนำข้อมูลที่ใช้ในการทดสอบ (Test data) มาทำการแบ่งกลุ่ม ขั้นตอนการตัดกิ่งนี้จะช่วยตัดกิ่งที่มีอัตราความผิดพลาดจากต้นไม้ โดยการเลือกต้นไม้ย่อย (Sub tree) ที่มีอัตราความผิดพลาดโดยประมาณน้อยที่สุดไว้ (ตัดต้นไม้ย่อยที่มีอัตราการผิดพลาดมากๆ ออกไป)

4.3 อัลกอริทึม SLIQ (SLIQ Algorithm)

SLIQ ใช้เทคนิคการเรียงข้อมูลก่อน (pre-sorting) ในการสร้างต้นไม้เพื่อลดเวลาในการประเมินหาค่าของการแบ่งกลุ่มของข้อมูลที่มีแอตทริบิวต์แบบตัวเลข (numerical attribute) โดยจะใช้วิธีการสร้างต้นไม้แบบเบรทเฟิส์ต (Breadth-first) ร่วมด้วย เพื่อทำให้สามารถแบ่งกลุ่มข้อมูลที่อยู่บนดิสก์ได้ ยิ่งไปกว่านั้น SLIQ ยังใช้วิธีการแบ่งเซ็ตย่อย (subsetting algorithm) สำหรับแอตทริบิวต์แบบตัวอักษร (categorical attribute) การนำเทคนิคต่างๆ ข้างต้นทำให้ SLIQ สามารถรองรับข้อมูลขนาดใหญ่ และการแบ่งกลุ่มที่มีกลุ่ม (classes), แอตทริบิวต์และข้อมูลจำนวนมาก ๆ ได้

4.3.1 การเรียงข้อมูลก่อน และการสร้างต้นไม้แบบ เบรทเฟิส์ต (Pre-Sorting and Breadth-First Growth)

เวลาที่ใช้ในการเรียงข้อมูลสำหรับแอตทริบิวต์แบบตัวเลขเป็นส่วนที่สำคัญในช่วงการค้นหาค่าของการแบ่งกลุ่มที่ดีที่สุด (best split) สำหรับแต่ละโหนดของต้นไม้ โดย SLIQ ลดเวลาในการเรียงข้อมูลลงด้วยการเรียงข้อมูลที่เป็นแอตทริบิวต์แบบตัวเลขเพียงครั้งเดียวในตอนเริ่มต้นการสร้างต้นไม้ โดยเรียกเทคนิคนี้ว่าพรีซอร์ทติ้ง (Pre-sorting)

ซึ่งก่อนการทำพรีซอร์ทติ้ง จะต้องทำการจัดโครงสร้างข้อมูลของตารางข้อมูลที่จะนำมาใช้ในการสร้างดิซิชั่นทรี โดยนำแอตทริบิวต์มาแยกออกเป็นลิสต์ ซึ่งมีลักษณะดังต่อไปนี้

- แอตทริบิวต์ลิส (Attribute List): แยกแต่ละแอตทริบิวต์ออกมาเป็นแต่ละแอตทริบิวต์ลิส ซึ่งจะประกอบด้วยฟิลด์ 2 ฟิลด์คือ ค่าของแอตทริบิวต์ และ ค่าดรรชนีที่อ้างอิงค่าที่ตรงกันใน Class List (index)
- คลาสลิส (Class List) สำหรับคลาสลาเบล (เป็นค่าของแอตทริบิวต์ที่จะนำมาใช้ในการทำนาย เช่น Churn กับ Retain) ซึ่งประกอบด้วยฟิลด์ 2 ฟิลด์คือ คลาสลาเบล และตัวอ้างอิงไปยังลีฟโหนดบนดิซิชั่นทรี (leaf reference)

โดยที่ลำดับของคลาสลิส จะเป็นลำดับเดียวกันกับลำดับของตารางข้อมูล ซึ่งลำดับแต่ละตัวเป็นตัวชี้ (Index) ที่ใช้ในการอ้างอิงถึงกันของแต่ละแอตทริบิวต์ลิส

จากรูปที่ 4.2 แสดงตารางซึ่งประกอบด้วยแอตทริบิวต์ดังนี้คือ Type of System (Sys_Type), Customer Age (Cust_Age), Subscriber Age (Subs_Age) และ Class โดยในแอตทริบิวต์ Sys_Type คือชนิดของระบบโทรศัพท์เคลื่อนที่ประกอบด้วยค่า A (Analog), D (Digital) และ P (Pre-paid) และแอตทริบิวต์ Cust_Age คืออายุของลูกค้า และแอตทริบิวต์ Subs_Age คือจำนวนปีที่ลูกค้าใช้บริการโทรศัพท์เคลื่อนที่ และกลุ่มที่จะแบ่ง (Class) มีค่าคือ

Training Data

	Sys_Type	Cust_Age	Subs_Age	Class
1	A	50	1	Y
2	P	19	2	N
3	D	23	1	Y
4	D	30	3	N
5	A	47	3	Y
6	D	43	2	Y
7	P	30	4	N
8	D	21	4	N
9	P	55	2	Y
10	D	24	4	N

After Pre-sorting

Sys_Type	Index
A	1
P	2
D	3
D	4
A	5
D	6
P	7
D	8
D	9
P	10

Type List

Cust_Age	Index
19	2
21	8
23	3
24	10
30	4
30	7
43	6
47	5
50	1
55	9

Cust_Age List

Subs_Age	Index
1	1
1	3
2	2
2	6
2	9
3	4
3	5
4	7
4	8
4	10

Subs_Age List

	Class	Leaf
1	Y	N1
2	N	N1
3	Y	N1
4	N	N1
5	Y	N1
6	Y	N1
7	N	N1
8	N	N1
9	Y	N1
10	N	N1

Class List

รูปที่ 4.2 ตัวอย่างโครงสร้างข้อมูล และการทำพรีซอร์ทติ้ง

"Y" = Churn และ "N" = Retain และจะนำแต่ละแอตทริบิวต์มาแยกออกเป็นแอตทริบิวต์ลิสแล้วจึงนำแอตทริบิวต์ลิสต์ที่เป็นแบบตัวเลขมาทำการเรียงลำดับจากน้อยไปหามากอย่างเป็นอิสระซึ่งกันและกัน

ลีฟโหนดของต้นไม้จะแสดงกลุ่มของข้อมูล และแต่ละกลุ่มจะเชื่อมกันด้วยเส้นทาง (Path) และโหนดไปยังรูทโหนดของต้นไม้ ดังนั้นคลาสลิสจะเป็นตัวบ่งชี้ว่าข้อมูลนั้นๆ อยู่ในกลุ่มใด โดยในตอนเริ่มต้นนั้น ฟิลด์ที่อ้างถึงลีฟ (Leaf reference) ทุกตัวในคลาสลิสจะกำหนดให้ชี้ไปยังรูทโหนดของต้นไม้

ขั้นตอนในการสร้างต้นไม้จะใช้วิธีการสร้างแบบเบรทเฟิร์ส คือแตกกิ่งจากบนลงล่าง โดยแตกกิ่งให้ครบจากซ้ายไปขวาของระดับ (Level) นั้น ๆ ก่อนจึงจะเริ่มแตกกิ่งลงไปยังระดับ (Level) ถัดไปเรื่อยๆ หลังจากการทำพรีซอร์ทติ้ง แอตทริบิวต์ลิสแต่ละตัว จะถูกนำมาหาค่าของการแบ่งกลุ่มที่ดีที่สุด (best split) ดังแสดงในรูปที่ 4.3

```
EvaluateSplits()

    for each attribute A do
        traverse attribute 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 ≤ 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 4.3 การหาค่าของการแบ่งกลุ่ม (Evaluating Splits)

4.3.2 การคำนวณหาค่าของการแบ่งกลุ่ม

โดยในการหาค่าที่ชี้การแบ่งกลุ่ม (Splitting index) สำหรับแต่ละแอตทริบิวต์ ใช้สูตร

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

ให้ T คือข้อมูลที่ใช้ในการเทรนนิ่ง (example)

p คือค่าความเป็นไปได้ของกลุ่ม j

$$best\ split = gini(T) - [((|S_1|/|S|) \times gini(S_1)) + ((|S_2|/|S|) \times gini(S_2))]$$

โดย S_1 และ S_2 เป็นกลุ่มที่แบ่งออกมาจากกลุ่ม S

ในการหาค่าของการแบ่งกลุ่มที่ดีที่สุด (Best split) บนลีฟโหนดจะมีสถิติความถี่ของกลุ่ม (Class histogram) ที่แสดงให้เห็นการกระจายความถี่ของค่าแต่ละค่าบนแอตทริบิวต์ โดยในแอตทริบิวต์แบบตัวเลข histogram เป็นลิสต์ซึ่งประกอบด้วย class (กลุ่มในการแบ่ง) และ frequency (ความถี่ของค่าในกลุ่มนั้นๆ) และในแอตทริบิวต์แบบหมวดหมู่ (categorical) นั้น สถิติความถี่ (Histogram) เป็นลิสต์ที่ประกอบด้วยค่าบนแอตทริบิวต์ (attribute value), กลุ่มในการแบ่ง (class), ความถี่ของค่าในกลุ่มนั้น ๆ (frequency)

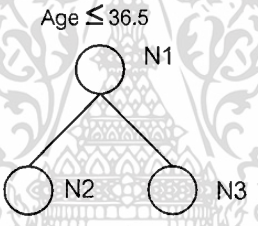
หลังจากคำนวณหาค่าของการแบ่งกลุ่มที่ดีที่สุด (best split) ของทุก ๆ ค่าในทุก ๆ แอตทริบิวต์ครบหมดแล้ว เปรียบเทียบค่าของการแบ่งกลุ่มที่ดีที่สุด (best split) ทั้งหมด ค่าที่สูงที่สุดจะถูกนำมาเป็นค่าของการแบ่งกลุ่มที่ดีที่สุด (best split)ของโหนดนั้นๆ ซึ่งรายละเอียดในการหาค่าของการแบ่งกลุ่มที่ดีที่สุด (best split) ซึ่งจำแนกตามประเภทของแอตทริบิวต์สามารถอธิบายได้ดังต่อไปนี้

- แอตทริบิวต์แบบตัวเลข (Numerical attribute)

  ใช้รูปแบบ $A \leq v$ ซึ่ง v เป็นจำนวนจริงของแอตทริบิวต์ A ในการแบ่งกลุ่มข้อมูลที่มีแอตทริบิวเป็นแบบตัวเลข โดยขั้นตอนแรกในการหาค่าของการแบ่งกลุ่มที่ดีที่สุด (best split) จะทำการเรียงข้อมูลตามค่าของแอตทริบิวต์ โดยให้ $v_1, v_2, ..., v_n$ เป็นค่าในแอตทริบิวต์ A ที่เรียงลำดับแล้ว

เมื่อนำค่าการแบ่งกลุ่มทุกค่าที่ได้มาเปรียบเทียบกัน สมมติว่าค่าสูงสุดที่หาได้คือ v_i ดังนั้น ค่าของการแบ่งกลุ่มที่ดีที่สุด (best split) ที่จะนำมาใช้ในการแบ่ง คือค่ากลางระหว่าง $v_i - v_{i+1}$ ตัวอย่างเช่น การคำนวณหาค่าของการแบ่งกลุ่มที่ดีที่สุด (best split) ของอายุลูกค้า Customer Age $\leq$ 30 ได้ค่าสูงสุดเมื่อเทียบกับแอตทริบิวต์เดียวกัน และแอตทริบิวต์อื่น และดังนั้นค่าของการแบ่งกลุ่มที่ดีที่สุด (best split) คือค่ากลางระหว่าง 30 และ 43 ซึ่งเท่ากับ 36.5 ดังนั้นค่าที่ใช้ในการแบ่งกลุ่มในชั้นแรกคือ Customer Age $\leq$ 36.5 จะได้ต้นไม้ดังรูปที่ 3.4 และค่าลีฟโหนด แต่ละตัวใน class list จะเปลี่ยนไปตามเงื่อนไขการแบ่งกลุ่มข้างต้น

เมื่อได้ต้นไม้ดังรูปที่ 4.4 แล้ว ขั้นต่อไปจะทำการพิจารณา N2 และ N3 ซึ่งแยกออกมาจาก N1 โดยพิจารณาเฉพาะค่าของแอตทริบิวต์ที่ถูกแบ่งกลุ่มตามเงื่อนไขในตอนแรกเท่านั้น และทำวนซ้ำวิธีเดิมจนกว่าในหนึ่งโหนดจะมีเพียงหนึ่งกลุ่ม (class) เท่านั้น



Class	Leaf
Y	N3
N	N2
Y	N2
N	N2
Y	N3
Y	N3
N	N2
N	N2
Y	N3
N	N2

รูปที่ 4.4 การแบ่งกลุ่มของต้นไม้ในระดับที่ 1

- แอตทริบิวต์แบบตัวอักษร (Categorical attribute)

ถ้า $S(A)$ เป็นเซ็ตของค่าที่เป็นไปได้ของ Attribute A (ซึ่งเป็นแอตทริบิวต์แบบ Categorical) โดย $A \in S'$ เมื่อ $S' \subset S$ โดยมีวิธีในการแบ่งกลุ่ม 2 แบบคือ

1) Greedy algorithm โดยจะทำการนำค่าจาก S ทีละค่าใส่ในกลุ่ม S' และคำนวณหาค่า Splitting index ของแต่ละเซ็ตย่อย และหลังจากนำค่าใน S ใส่ลงใน S' จนครบทุกตัวแล้ว ให้นำค่าที่ชี้การแบ่งกลุ่ม (splitting index) ของแต่ละเซ็ตย่อยมาเปรียบเทียบหาค่า best split สำหรับแอตทริบิวต์นั้นๆ Hybrid algorithm หาค่าการแบ่งกลุ่มจากทุกความเป็นไปได้ในการแบ่งเซ็ตย่อย และนำแต่ละเซ็ตย่อยมาหาค่าที่ชี้การแบ่งกลุ่ม (splitting index) และนำค่า ค่าที่ชี้การแบ่งกลุ่ม (splitting index) ของแต่ละเซ็ตย่อยมาเปรียบเทียบหาค่าที่ชี้การแบ่งกลุ่ม (splitting index) สำหรับแอตทริบิวต์นั้นๆ ซึ่งจำนวนของเซ็ตย่อย (subset) มีค่าเท่ากับ 2^n เซ็ต สำหรับแอตทริบิวต์ที่มีค่าที่เป็นไปได้ n ค่า โดยจะต้องมีการกำหนดค่า MAXSETSIZE คือจำนวน n สูงสุดที่ทำให้การหาค่า split มีประสิทธิภาพ โดยทั่วไปจะกำหนด MAXSETSIZE = 10
2) ตัวอย่างเช่น แอตทริบิวต์ Sys_Type ซึ่งเป็นแอตทริบิวต์แบบตัวอักษร ที่มีสมาชิกคือ {A, P, D} สามารถแบ่งเป็นเซ็ตย่อยตามหลัก Greedy algorithmได้ดังนี้คือ {{A}, {P, D}}, {{A, P}, {D}} และ {{A, D}, {P}} และคำนวณหาค่าที่ชี้การแบ่งกลุ่ม (splitting index) ของแต่ละเซ็ตย่อย โดยใช้ Histogram ดังรูปที่ 4.5 ซึ่งประกอบด้วย <attribute value, class, frequency> มาช่วยในการคำนวณ

	Y	N
A	2	0
P	0	3
D	3	2

รูปที่ 4.5 สถิติความถี่สำหรับแอตทริบิวต์แบบตัวอักษร

4.3.3 การปรับปรุงแก้ไขคลาสลิส

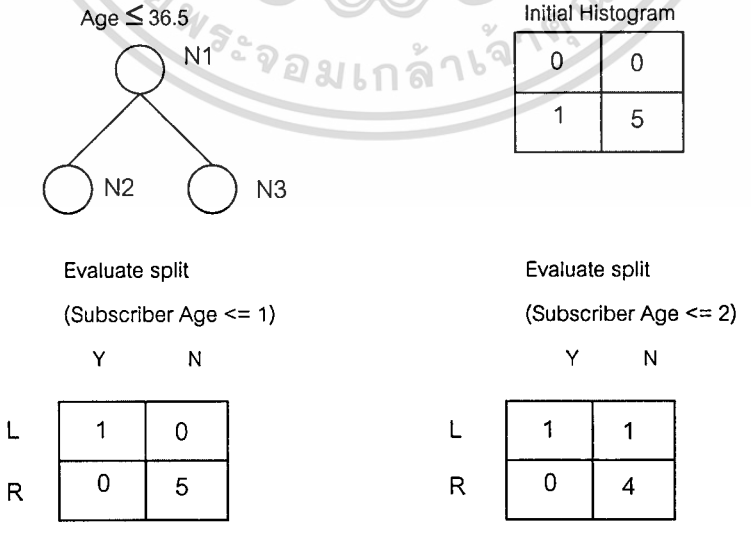
หลังจากสร้างโหนดลูก (Child node) สำหรับแต่ละลีฟโหนดแล้ว จะต้องทำการปรับปรุงแก้ไขฟิลด์ที่อ้างถึงลีฟ (Leaf reference) ในคลาสลิสตามขบวนการดังรูปที่ 4.6

```
UpdateLabels()
    for each attribute A used in a split do
        traverse attribute list of A
        for each value v in the attribute list do
            find the corresponding entry in the class list (say 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 4.6 อัลกอริทึมปรับปรุงแก้ไขคลาสลิส (Updating Class List Algorithm)

4.3.4 ตัวอย่างในการสร้างต้นไม้

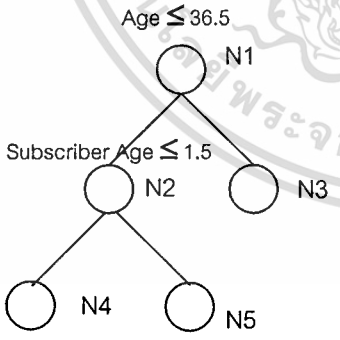
จากต้นไม้ดังรูปที่ 4.7 พิจารณาคลาสลาเบลของแต่ละโหนด จะเห็นว่าโหนด N3 นั้น มีคลาส-ลาเบลเป็น Y เพียงตัวเดียวเท่านั้น ดังนั้นจึงไม่ต้องทำการแบ่งต่อแล้ว



รูปที่ 4.7 การหาค่าการแบ่งกลุ่มของต้นไม้ในระดับที่ 2

ในโหนด N2 คลาสลาเบลยังคงมีทั้งค่า Y และ N ดังนั้นต้องแตกโหนด โดยจะพิจารณาที่แอตทริบิวต์อีกสองแอตทริบิวต์ที่เหลือ คือ subscriber age และ type of system โดยหาค่าการแบ่งกลุ่มจากค่าบนแอตทริบิวต์ที่ถูกแยกโดยเงื่อนไข Customer Age $\leq$ 36.5 จากรูปที่ 4.7 แสดงตัวอย่างในการพิจารณาค่าในการแบ่งกลุ่มบนแอตทริบิวต์ subscriber age โดยพิจารณาร่วมกับสถิติความถี่ (Histogram) ในการหาค่าการแบ่งกลุ่มทุกครั้งจะใช้สถิติความถี่ของกลุ่ม (Class histogram) ในการพิจารณาหาค่าด้วย) โดยในสถิติความถี่ของกลุ่ม (Class histogram) ค่า L จะแสดงค่าการแบ่งข้อมูลตามเงื่อนไขเป็นจริง และค่า R แสดงค่าการแบ่งข้อมูลตามเงื่อนไขเป็นเท็จจากรูปที่ 4.7 สถิติความถี่เริ่มต้น (Initial histogram) ค่า L จะมีค่า Y = 0 และ N = 0 เนื่องจากเป็นสถิติความถี่ (Histogram) เริ่มต้นก่อนทำการแบ่งกลุ่ม จึงยังไม่มีค่าตามเงื่อนไขที่กำหนด ซึ่งค่าทั้งหมดถูกเก็บไว้ใน R โดยแบ่งเป็น Y = 1 และ N = 5 (จำนวนของค่าบนแอตทริบิวต์ subscriber age หลังจากการแบ่งในระดับแรกจำนวนทั้งหมด 6 ตัว)

โดยในการพิจารณาที่ค่าบนแอตทริบิวต์ที่ Subscriber Age $\leq$ 1 จะเห็นว่าค่าใน histogram เปลี่ยนแปลงไป โดย ค่า L มีค่า Y = 1 กล่าวคือ ถ้าทดลองแบ่งตามเงื่อนไขข้างต้นจะมีค่าบนแอตทริบิวต์ที่มี class label เท่ากับ Y จำนวนหนึ่งค่า และ N = 0 คือไม่มีค่าบนแอตทริบิวต์ที่มีคลาสลาเบลเท่ากับ N และค่า R ที่เป็น Y จะมีค่าลดลงเป็นศูนย์ เนื่องจากมีการเปลี่ยนแปลงตามเงื่อนไข และย้ายค่าไปที่ L ที่เป็น Y และเมื่อพิจารณาในเงื่อนไขถัดไป ค่าในสถิติความถี่ของกลุ่ม (Class histogram) จะเปลี่ยนแปลงไปตามเงื่อนไขที่ใช้ในการพิจารณาครั้งนั้นๆ



Class	Leaf
Y	N3
N	N5
Y	N4
N	N5
Y	N3
Y	N3
N	N5
N	N5
Y	N3
N	N5

รูปที่ 4.8 ตัวอย่างต้นไม้ที่เสร็จสมบูรณ์แล้ว

หลังจากทำการพิจารณาหาค่าการแบ่งกลุ่มที่ดีที่สุดแล้วจะได้ต้นไม้ดังรูปที่ 4.8 โดยค่าที่นำมาใช้ในการแบ่งกลุ่มในโหนด N2 คือ subscriber age ≤ 1.5 และค่าของลีฟโหนดบนคลาสลิสจะเปลี่ยนแปลงค่าต่างๆ ตามการแบ่งกลุ่มที่เปลี่ยนแปลงไป จากรูปที่ 4.8 จะเห็นว่าทุกลีฟโหนดมีกลุ่มอยู่เพียงหนึ่งกลุ่มเท่านั้น ดังนั้นจึงไม่ต้องทำการแตกโหนดต่อไปแล้ว โหนดบางโหนดอาจเป็นโหนดที่มีสมาชิกทุกตัวอยู่ในกลุ่มเดียวกันได้เร็วกว่าโหนดอื่นๆ ซึ่งช่วยให้ย่อขนาดของแอตทริบิวต์ลิสต์ในการนำมาพิจารณาแบ่งกลุ่มเล็กลง ทำให้มีการทำงานเร็วขึ้น

4.3.5 การตัดกิ่ง (Pruning)

ในการตัดกิ่งเป็นการประมาณค่าความถูกต้อง หรือค่าความผิดพลาดจากต้นไม้ที่สร้างมาจาก – ข้อมูลที่ใช้สร้างโมเดล (Training data) และทำการตัดกิ่งต้นไม้ เพื่อทำให้ต้นไม้มีค่าความผิดพลาดน้อยที่สุด โดยสามารถแบ่งออกได้เป็นสองแบบ คือ การตัดกิ่งก่อนสร้างเสร็จ (Pre-Pruning) และการตัดกิ่งหลังสร้างเสร็จ (Post-Pruning)

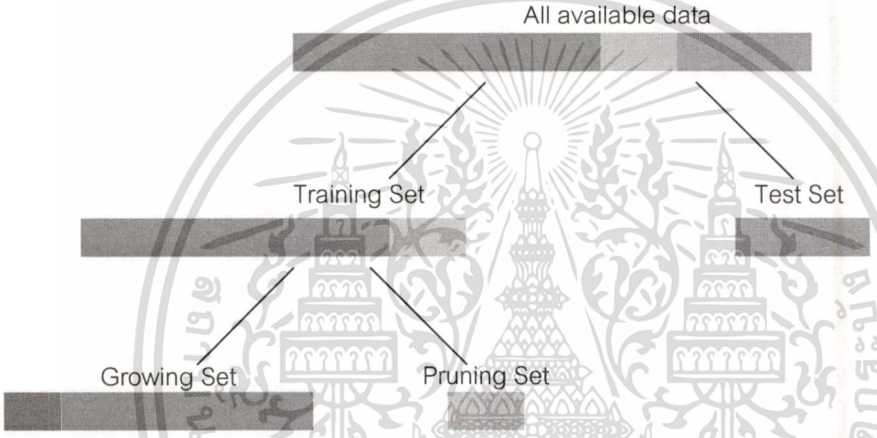
การตัดกิ่งก่อนสร้างเสร็จ (Pre-Pruning) นั้นเป็นวิธีการตัดกิ่ง (Pruning) ในช่วงการสร้างต้นไม้ คือก่อนที่จะทำการแตกกิ่ง จะทำการคำนวณค่าความถูกต้อง (Accuracy) ของการมีลูก หรือการแตกกิ่งต่อ กับการไม่มีลูก หรือไม่ทำการแตกกิ่งต่อ ว่าแบบใดมีค่าความถูกต้องมากกว่ากัน ถ้าการแตกกิ่งต่อมีค่าความถูกต้องมากกว่าก็จะทำการแตกกิ่งต่อไป แต่กลับกันถ้ามีค่าความถูกต้องน้อยกว่าก็จะไม่ทำการแตกกิ่งต่อ และจะถือว่าโหนดนั้นเป็นลีฟโหนด โดยการคำนวณหาค่าความถูกต้องจะนับจำนวนของข้อมูลที่โมเดลทำการทำนายถูกต้องทั้งหมด มาหารด้วยจำนวนของข้อมูลในส่วนที่ใช้การตัดกิ่ง (Pruning) ทั้งหมด

การตัดกิ่งหลังสร้างเสร็จ (Post-Pruning) เป็นวิธีการตัดกิ่ง (Pruning) หลังจากขั้นตอนการสร้างต้นไม้เสร็จสิ้นแล้ว (เมื่อทุกลีฟโหนดเป็นกลุ่มของลาเบลที่สมาชิกทุกตัวมีค่าเดียวกันหมด) จะทำการประเมินค่าความถูกต้องของ รูทโหนดของแต่ละต้นไม้ย่อย (Subtree) และโหนดของต้นไม้ย่อย (Subtree) นั้นๆ ถ้าค่าความถูกต้องของรูทโหนดมีมากกว่าก็จะทำการตัดโหนดของต้นไม้ย่อย (Subtree) และให้รูทของต้นไม้ย่อย (Subtree) นั้นเป็นลีฟ และหยุดการตัดกิ่ง (Pruning) เมื่อค่าความถูกต้องเปลี่ยนแปลงมากเกินไป

โดยทั่วไปมีวิธีการในการประเมินค่าความถูกต้องสองวิธีใหญ่ๆ โดยวิธีแรกจะใช้ข้อมูลที่ใช้สร้างโมเดล (Training data) เดิมมาใช้ในการประเมินค่าผิดพลาด วิธีนี้เรียกว่า cross-validation โดยจะนำข้อมูลจากข้อมูลที่ใช้สร้างโมเดล (Training data) มาแบ่งเป็นหลายๆ ตัวอย่าง และสร้างต้นไม้จากแต่ละตัวอย่างนั้นๆ และนำต้นไม้หลายๆ ต้นที่ได้มาใช้ในการประเมินค่าผิดพลาดของต้นไม้

ย่อยของต้นไม้ที่ได้ในตอนแรกสุด และเลือกต้นไม้ที่กระชับซึ่งให้ค่าความถูกต้องสูง ซึ่งวิธีนี้จะไม่เหมาะกับเซ็ตข้อมูลขนาดใหญ่ เพราะจะทำให้สิ้นเปลืองค่าใช้จ่ายในการสร้างต้นไม้

และวิธีที่สองจะแบ่งข้อมูลที่ใช้สร้างโมเดล (Training data) เป็นสองส่วน ส่วนหนึ่งใช้ในการสร้างต้นไม้ เรียกว่ากลุ่มข้อมูลที่ใช้สร้าง (Growing set) และอีกส่วนหนึ่งใช้ในการตัดกิ่ง (Pruning) เรียกว่ากลุ่มข้อมูลที่ใช้ตัดกิ่ง (Pruning Set) ดังรูปที่ 4.9 ซึ่งแสดงสัดส่วนการแบ่งข้อมูลจากข้อมูลที่จะใช้ในการสร้างโมเดล และนำข้อมูลในกลุ่มข้อมูลที่ใช้ตัดกิ่ง (Pruning Set) มาทำการประมาณค่าความถูกต้อง (accuracy) ของต้นไม้ และเลือกต้นไม้ย่อย (Subtree) ที่มีค่าความถูกต้องมากกว่าไว้



รูปที่ 4.9 สัดส่วนการแบ่งข้อมูลจากข้อมูลที่จะใช้ในการสร้างโมเดล

การแบ่งสัดส่วนข้อมูลนั้นควรที่จะเลือกสัดส่วนที่เหมาะสม มิเช่นนั้นแล้วอาจนำไปสู่ความผิดพลาดที่มากขึ้นได้ และในการใช้ข้อมูลในส่วนของกลุ่มข้อมูลที่ใช้ตัดกิ่ง (Pruning Set) มาใช้ในการตัดกิ่ง (Pruning) เพียงอย่างเดียวนั้น ทำให้จำนวนข้อมูลในส่วนของกลุ่มข้อมูลที่ใช้สร้าง (Growing set) ที่จะนำมาใช้ในการสร้างโมเดลลดลง ซึ่งจะทำให้ต้นไม้มีความถูกต้องลดน้อยลง ดังนั้นในการเลือกวิธีการตัดกิ่ง (Pruning) ที่ดีนั้น ต้องพิจารณาทั้งข้อดี และข้อเสีย เพื่อให้ได้วิธีการที่เร็ว ได้ต้นไม้ที่กระชับ และถูกต้อง

4.4 การวัดค่าความถูกต้อง (Accuracy)

หลังจากเสร็จสิ้นขั้นตอนการตัดกิ่ง (Pruning) จนได้ต้นไม้ที่กระชับ และถูกต้องแล้ว สิ่งที่ต้องทำในถัดมาคือ การวัดความแม่นยำ หรือความถูกต้อง (Accuracy) ของต้นไม้ และแสดงถึง

ประสิทธิภาพ (Performance) ของ เทคนิคการบ่งกลุ่ม (Classification) โดยใช้ข้อมูลในส่วนของกลุ่มข้อมูลที่ใช้ทดสอบ (Test set) ทำการวัดความสามารถในการแบ่งกลุ่มได้ถูกต้องของโมเดลโดยไม่สนใจค่า หรือลาเบลที่แท้จริงในกลุ่มข้อมูลที่ใช้ทดสอบ (Test set) โดยโมเดลจะพยายามทำนายลาเบลนั้นๆ แทน ตัวอย่างการวัดค่าความถูกต้องเช่น ถ้าโมเดลทำนายว่า “Y” หรือ “N” สำหรับข้อมูล 100 เรคคอร์ด และทำนายถูกเป็นจำนวน 80 เรคคอร์ด แล้วค่าความถูกต้องของโมเดลคือ 80 เปอร์เซ็นต์

การวัดค่าความถูกต้องของโมเดลอาจไม่เหมาะกับบางสถานการณ์ เมื่อเหตุการณ์ที่สนใจมีจำนวนน้อยเกินไป เช่นโมเดลการทำนายการยกเลิกบริการของลูกค้ามีค่าความถูกต้อง 98 เปอร์เซ็นต์ จากค่าความถูกต้องดังกล่าวดูเสมือนว่าโมเดลนี้มีค่าความถูกต้องที่น่าเชื่อถือ แต่จะไม่เป็นเช่นนั้นเมื่อมีเพียง 2 เปอร์เซ็นต์ที่เป็นลูกค้าที่มียกเลิกบริการ

บทที่ 5
ระบบการวิเคราะห์หาสาเหตุการยกเลิกบริการโทรศัพท์เคลื่อนที่

ในบทนี้เป็นรายละเอียดทั้งหมดของระบบดาต้าไมนิ่งในการวิเคราะห์หาสาเหตุการยกเลิกบริการโทรศัพท์เคลื่อนที่โดยใช้ดิซิชั่นทรี โดยระบบจะรับข้อมูลที่จะใช้ในการสร้างโมเดลเป็นอินพุตท์ และระบบจะทำการประมวลผลข้อมูลด้วยกระบวนการทางดาต้าไมนิ่ง เพื่อให้ได้เอาพุตท์ออกมาเป็นโมเดลในแบบดิซิชั่นทรี ซึ่งในบทนี้จะกล่าวถึงโครงสร้างการทำงานของระบบ และรายละเอียดการทำงานของทั้งระบบ

5.1 เครื่องมือที่ใช้ในการพัฒนาระบบ

ระบบการวิเคราะห์หาสาเหตุการยกเลิกบริการโทรศัพท์เคลื่อนที่นี้ ใช้ JBuilder ในการพัฒนาระบบ ซึ่ง JBuilder เป็นเครื่องมือการพัฒนาระบบด้วยภาษาจาวา (Java Language) เหตุผลที่เลือกใช้ภาษาจาวาด้วย JBuilder ในการพัฒนาระบบนี้ เนื่องจากภาษาจาวาเป็นภาษาเชิงวัตถุ (Java Programming Language) ซึ่งให้ความยืดหยุ่น, ความเป็นมาตรฐาน, ความชัดเจน และกลไกซึ่งส่งเสริมการนำโปรแกรมที่สร้างไว้แล้วมาใช้งานใหม่ได้ (Reusability) ทำให้สามารถนำบาง method และ class ที่มีส่วนเกี่ยวข้องกับระบบซึ่งมีอยู่แล้ว และมาใช้ในการพัฒนาระบบได้ รวมทั้ง method และ class ของระบบยังสามารถนำไปใช้ หรือนำไปพัฒนากับระบบอื่น หรือพัฒนาระบบเดิมให้ดีขึ้นได้ และภาษาจาวามีส่วนของ Swing Component ซึ่งทำให้หน้าจอการใช้งาน (User Interface) มีลักษณะเหมือนกับโปรแกรมทั่วไปบนวินโดว์ที่ผู้ใช้คุ้นเคย และใช้งานได้ง่าย และสามารถพัฒนาระบบให้ติดต่อกันฐานข้อมูลผ่าน JDBC API ทำให้ระบบสามารถเข้าถึง (Access) และ ปรับปรุงแก้ไข (Manipulating) ฐานข้อมูลต่างๆ ได้เช่น Microsoft Access, Oracle เป็นต้น ซึ่งทำให้การพัฒนาระบบนี้สามารถติดต่อกับฐานข้อมูลได้หลากหลาย เพื่อนำข้อมูลมาใช้ในการสร้างโมเดลต่อไป

5.2 แนวทางในการพัฒนาระบบ

ระบบนี้พัฒนาให้อยู่ในรูปแบบของแอปพลิเคชั่นบนวินโดวส์ ซึ่งเป็นระบบปฏิบัติการที่นิยมใช้กันอย่างแพร่หลาย เพื่อให้ผู้ใช้สามารถเรียนรู้ และใช้งานระบบได้ง่าย เนื่องจากผู้ใช้ส่วนใหญ่มี

ความคุ้นเคยกับการใช้โปรแกรมในรูปแบบแอปพลิเคชั่นบนวินโดวส์อยู่แล้ว และเพื่ออำนวยความสะดวกในการเลือกข้อมูลจากฐานข้อมูลที่หลากหลายได้ จึงได้ทำการออกแบบให้ผู้ใช้สามารถเลือกติดต่อกับฐานข้อมูลที่ต้องการได้ ยิ่งไปกว่านั้นหากข้อมูลที่เลือกมามีค่าบางค่าขาดหายไป (Missing Value) ระบบจะมีส่วนแนะนำ และให้ผู้ใช้สามารถแทนที่ค่าที่ขาดหายไป หรือข้อมูลส่วนนั้นทิ้งไป ซึ่งจะส่งผลให้ได้ข้อมูลที่สมบูรณ์ เพื่อนำไปสร้างโมเดลที่มีประสิทธิภาพต่อไปได้ และในส่วนของการแสดงผลนั้น ได้มีการออกแบบให้ระบบแสดงผลโมเดลทั้งในแบบกราฟฟิก ซึ่งแสดงในรูปแบบของ Tree Graph ซึ่งมีลักษณะเหมือนกับการแสดงโฟลเดอร์ในวินโดวส์เอ็กซ์โพลเลอร์ และแสดงผลในลักษณะของ rules ซึ่งเข้าใจได้ง่ายอีกด้วย นอกจากนี้ผล (Output) ที่ได้สามารถบันทึกและนำกลับมาเปิดดูได้ในภายหลังด้วย ดังนั้นผู้ใช้สามารถเปิดดูโมเดลที่เคยสร้างไว้แล้ว โดยไม่จำเป็นต้องทำการสร้างโมเดลใหม่ทุกครั้งที่ต้องการ

5.3 โครงสร้างการทำงานของระบบ

ระบบประกอบด้วยส่วนของการรับ (Input) ข้อมูล, ส่วนของการเตรียมข้อมูล (Data Preprocessing), ส่วนการสร้างโมเดล และส่วนของการแสดงผล (Output) โดยแต่ละส่วนมีรายละเอียดดังต่อไปนี้

5.3.1 การรับข้อมูล (Input)

สามารถรับข้อมูลโดยการเลือกติดต่อกับฐานข้อมูลที่ต้องการ โดยระบุ URL และ Driver ของฐานข้อมูล รวมทั้ง User Name และ Password (ถ้ามี) หลังจากติดต่อกับฐานข้อมูลเรียบร้อยแล้ว จะทำการเลือกข้อมูลจากฐานข้อมูลโดยใช้คำสั่ง SQL ระบุแอตทริบิวต์ และตารางที่ต้องการ โดยหลักการในการเลือกข้อมูลนั้นควรเลือกข้อมูลที่เป็นตัวอย่างที่มีค่าลาเบลที่เป็นส่วนของลูกค้าที่ยกเลิก (Churn) และส่วนของลูกค้าที่ยังใช้บริการกับระบบอยู่ (Retain) เป็นจำนวนใกล้เคียงกัน และเลือกแอตทริบิวต์ที่จะนำมาใช้ในการสร้างโมเดลที่เกี่ยวข้อง และเหมาะสมกับการทำนายดังที่ได้กล่าวไว้แล้วในบทที่ 4 เพื่อให้ผลการสร้างโมเดลมีค่าความถูกต้องที่ดี และยอมรับได้

5.3.2 การเตรียมข้อมูล (Data Preparation)

เมื่อติดต่อ และเลือกข้อมูลจากฐานข้อมูลเรียบร้อยแล้ว ระบบจะแสดงรายการแอตทริบิวต์ทั้งหมดที่เลือกไว้ โดยแสดงชนิดของแอตทริบิวต์ และสถานะ (Status) แสดงถึงค่าที่สมบูรณ์ของแต่ละแอตทริบิวต์ โดยถ้ามีแอตทริบิวต์ที่มีค่าขาดหายไป (Missing Value) ระบบจะแสดงสถานะบอกว่าแอตทริบิวต์นั้นๆ ยังไม่สมบูรณ์มีข้อมูลบางส่วนขาดหายไป ซึ่งผู้ใช้จะต้องทำการแก้ไขค่าที่ขาดหายไป ด้วยการลบข้อมูลส่วนนั้นทิ้งไป หรือแทนที่ค่าข้อมูลส่วนนั้น โดยระบบจะมีค่าเฉลี่ยที่คำนวณจากข้อมูลที่มีอยู่ช่วยสนับสนุนการตัดสินใจในการแทนที่ค่าให้กับผู้ใช้

5.3.3 การสร้างโมเดล (Decision Tree Classifier)

หลังจากเตรียมข้อมูลเรียบร้อยแล้ว ในการสร้างโมเดลจะต้องทำการกำหนดค่าต่างๆ ดังนี้

- แอตทริบิวต์ที่สนใจ (Target Attribute) คือแอตทริบิวต์ที่มีค่าเป็นลาเบลที่มีการแบ่งกลุ่มไว้ล่วงหน้าแล้ว หรือกล่าวได้อีกนัยหนึ่งว่า เป็นแอตทริบิวต์ที่เราสนใจที่จะใช้ในการทำนายค่า สำหรับแอตทริบิวต์ที่ระบบนี้สนใจคือ ยกเลิกบริการ (Churn) หรือ ยังคงใช้บริการต่อไป (Retain) ซึ่งจะแทนด้วยลาเบล “Y” และ “N” โดย “Y” คือลูกค้ายกเลิกบริการ และ “N” คือลูกค้ายังคงใช้บริการต่อไป
- จำนวนสมาชิกต่ำสุดที่มีเหลืออยู่ในลีฟโหนด (Minimum Member in Leaf Node) คือเมื่อโหนดใดๆ มีจำนวนสมาชิกในโหนดเป็นจำนวนน้อยกว่าหรือเท่ากับที่ได้กำหนดไว้ จะหยุดการแตกกิ่งต่อ และให้โหนดนั้นเป็นลีฟโหนด
- ระดับสูงสุดของต้นไม้ (Maximum Tree Level) คือในการสร้างโมเดล หรือต้นไม้ (Tree) หากต้นไม้มีระดับ (Level) เท่ากับระดับสูงสุดของต้นไม้ที่กำหนดไว้ จะทำการหยุดสร้างต้นไม้ไว้ที่ระดับนั้น
- จำนวนข้อมูลที่จะใช้ในการสร้างต้นไม้ (Percentage of Training Set) คือขนาดของข้อมูลที่จะใช้ในการสร้างต้นไม้เทียบเป็นเปอร์เซ็นต์ โดยส่วนแรกจะเป็นข้อมูลที่ใช้สร้างโมเดล (Training data) และส่วนที่เหลือจะเป็นส่วนของกลุ่มข้อมูลที่ใช้ทดสอบ (Test set)

ในการกำหนดค่าจำนวนต่ำสุดของสมาชิกในลีฟโหนด และระดับของต้นไม้สูงสุดนั้นทำเพื่อป้องกันไม่ให้ต้นไม้มีขนาดใหญ่มากเกินไป หรือมีจำนวนโหนดมากเกินไป เพราะต้นไม้ขนาดใหญ่อาจทำให้ค่าการทำนายผิดพลาดมากขึ้น ดังนั้นหากการกำหนดค่าดังกล่าวจะช่วยประหยัดเวลาในการสร้างต้นไม้ และเวลาในการตัดกิ่ง (Pruning) ช่วยให้ระบบได้ผลลัพธ์เร็วขึ้น มี Response Time ที่ดี โดยการกำหนดค่าที่เหมาะสมนั้นจะต้องพิจารณาขนาดของข้อมูลทั้งหมดที่เลือกมา และค่าความถูกต้องที่ยอมรับได้ด้วย

5.3.4 การแสดงผล (Output)

ผล (Output) ที่ได้จากระบบ จะทำการแสดงผลใน 2 ส่วน โดยส่วนแรกเป็นการแสดงผลแบบกราฟฟิกในรูปแบบของ Tree Graph ที่เป็นลักษณะของโฟลเดอร์ โดยโหนดพ่อ (Parent Node) และโหนดลูก (Leaf Node) จะแสดงด้วยไอคอนที่แตกต่างกัน และแต่ละลีฟโหนดจะแสดงค่าลาเบลที่โมเดลทำนาย และแสดงเปอร์เซ็นต์ของจำนวนสมาชิกที่มีอยู่ในลีฟโหนดนั้นๆ เช่น ถ้าลีฟโหนดมีสมาชิกทั้งหมดจำนวน 10 ตัว เป็นลาเบล “Y” จำนวน 8 ตัว และเป็น “N” จำนวน 2 ตัว จะแสดงเป็นเปอร์เซ็นต์ดังนี้คือ Churn: 80% และ Retain: 20%

ส่วนที่โหนดพ่อ (Parent Node) จะแสดงเงื่อนไขในการแตกกิ่ง โดยข้อมูลที่ถูกเงื่อนไข (True) คือโหนดลูกทางซ้าย และข้อมูลที่ผิดเงื่อนไข (False) คือโหนดลูกทางขวา ตัวอย่างของรูปแบบการแสดงเงื่อนไขแบ่งตามชนิดของแอตทริบิวต์ที่เป็นแบบตัวเลข (Numerical) และ ตัวอักษร (Categorical) เช่น

- แอตทริบิวต์แบบตัวเลข (Numerical Attribute) โหนดพ่อแสดงเงื่อนไข "Age ≤ 30" แสดงว่าสมาชิกในโหนดลูกทางซ้ายจะเป็นสมาชิกที่มีอายุน้อยกว่า หรือเท่ากับ 30 ส่วนสมาชิกในโหนดลูกทางขวาจะเป็นสมาชิกที่มีอายุมากกว่า 30
- แอตทริบิวต์แบบตัวอักษร (Categorical Attribute) โหนดพ่อมีเงื่อนไข "Areas = {N, C, S}" แสดงว่าสมาชิกในโหนดลูกทางซ้ายจะเป็นสมาชิกที่มีค่า Areas เป็น 'N' หรือ 'C' หรือ 'S' และสมาชิกในโหนดของลูกทางขวาคือสมาชิกที่มีค่า Areasที่เหลือในแอตทริบิวต์นั้นนอกเหนือจากที่แสดงในเงื่อนไข

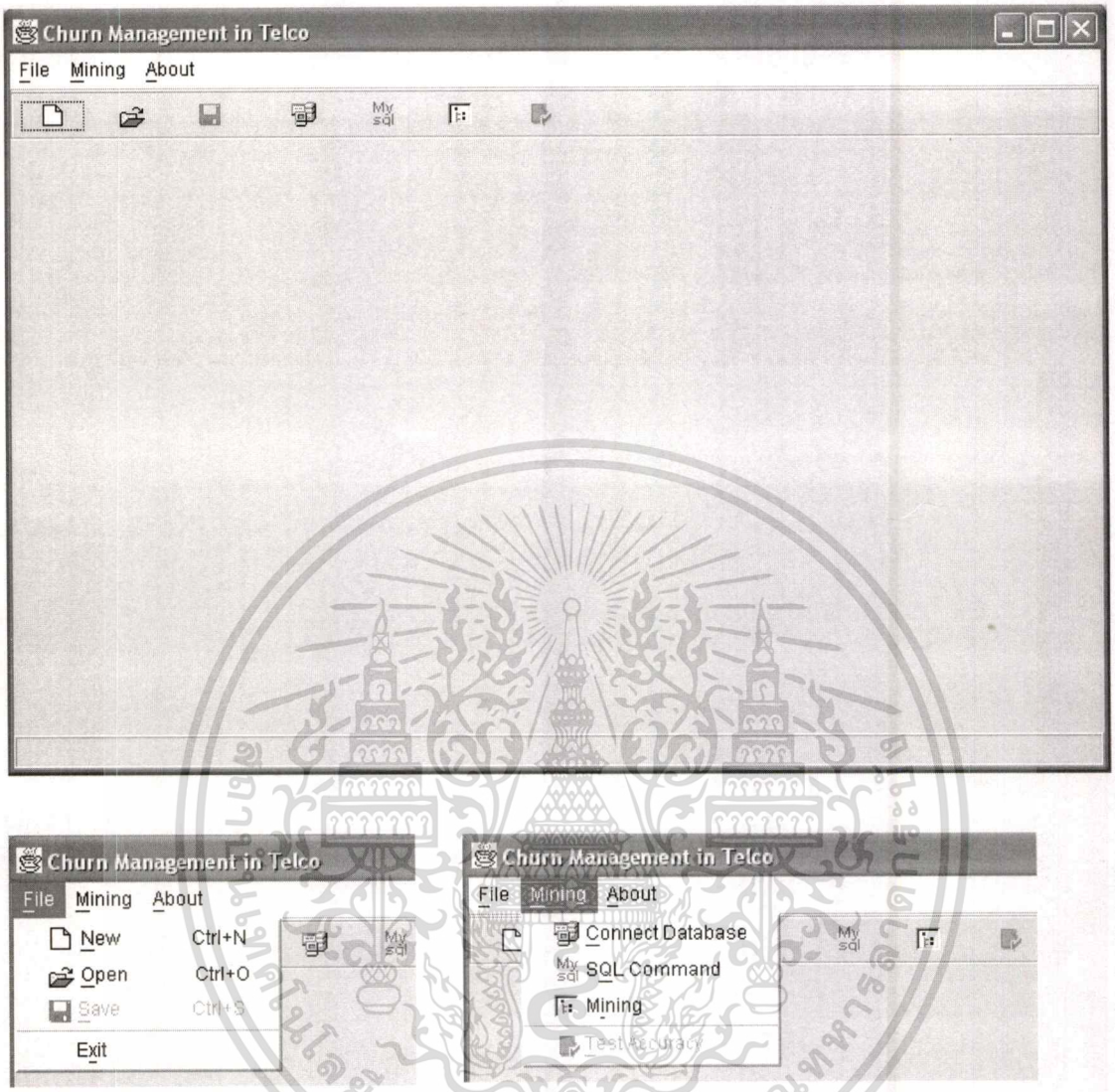
และในส่วนที่สองเป็นส่วนการแสดงผลของ Rule ที่แปลงมาจากโมเดลของต้นไม้ที่ได้ข้างต้น โดยมีหลักการคือ เริ่มจาก Root Nod e ถ้าค่าของแอตทริบิวต์ถูกต้องตามเงื่อนไขให้ไปทางซ้าย ถ้าผิดเงื่อนไขให้ไปทางขวา และท่องไปตามเงื่อนไขของโหนดต่อไปเรื่อยๆ จนถึงลีฟโหนด ถือเป็นหนึ่งเส้นทาง (Path) โดยจะแสดงเฉพาะ Rule ที่มีค่าลาเบลในลีฟโหนดเป็น "Y" คือ ลูกค้ายกเลิกบริการ (Churn) โดยแต่ละเส้นทาง (Path) ของโมเดลต้นไม้ (Decision Tree Model) เพื่อนำไปพัฒนาข้อบังคับทางธุรกิจ (Business Rule) ว่าลูกค้ากลุ่มไหนที่จะรวม หรือไม่รวมไว้ในแคมเปญ (Campaign) การตลาด

โดยผลที่ได้สามารถบันทึก (Save) เป็นแฟ้ม (File) ในรูปแบบของ *.txt และสามารถนำแฟ้มที่บันทึกโมเดลไว้มาเปิดดูโมเดลได้ในรูปแบบ Tree Graph และ Rule ได้เหมือนเดิม

5.4 รายละเอียดของหน้าจอการทำงาน

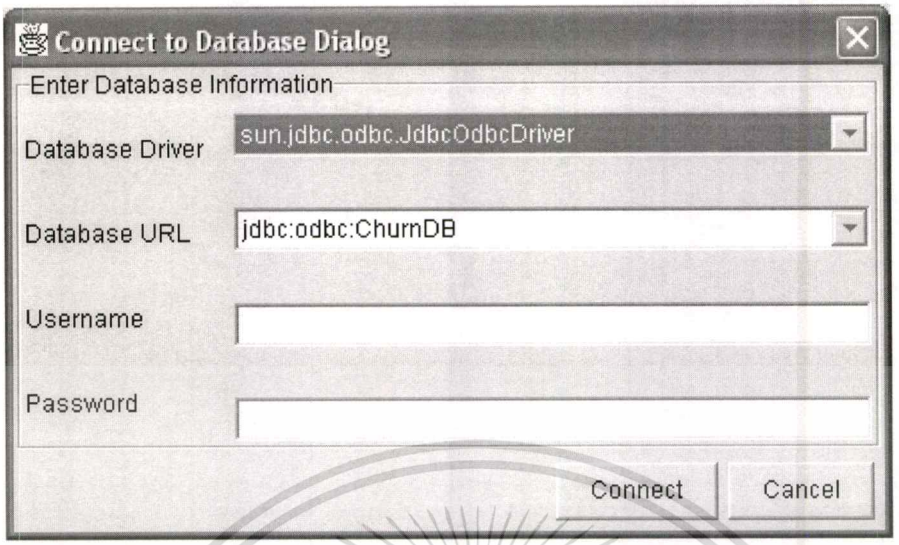
หน้าจอการทำงานของระบบการวิเคราะห์หาสาเหตุการยกเลิกบริการโทรศัพท์เคลื่อนที่ประกอบด้วยหน้าจอดังนี้

- หน้าจอแรก เป็นหน้าจอหลักของระบบ ดังแสดงในรูปที่ 5.1 โดยสามารถเลือกได้ว่าจะทำการสร้างโมเดลใหม่ หรือจะดูผลของโมเดลที่เคยสร้างไว้แล้ว ในการเลือกสร้างโมเดลใหม่สามารถทำได้สองแบบ คือเลือกเมนู หรือปุ่ม New ซึ่งจะขึ้นหน้าจอการติดต่อกับฐานข้อมูลและหน้าจอให้ใส่คำสั่ง SQL ตามลำดับให้โดยอัตโนมัติ หรืออาจเลือกสร้างโมเดลใหม่โดยเลือกเมนู หรือปุ่ม Connect Database เพื่อติดต่อกับฐานข้อมูล และหลังจากนั้นเลือกปุ่ม SQL Command ด้วยตัวเองก็ได้



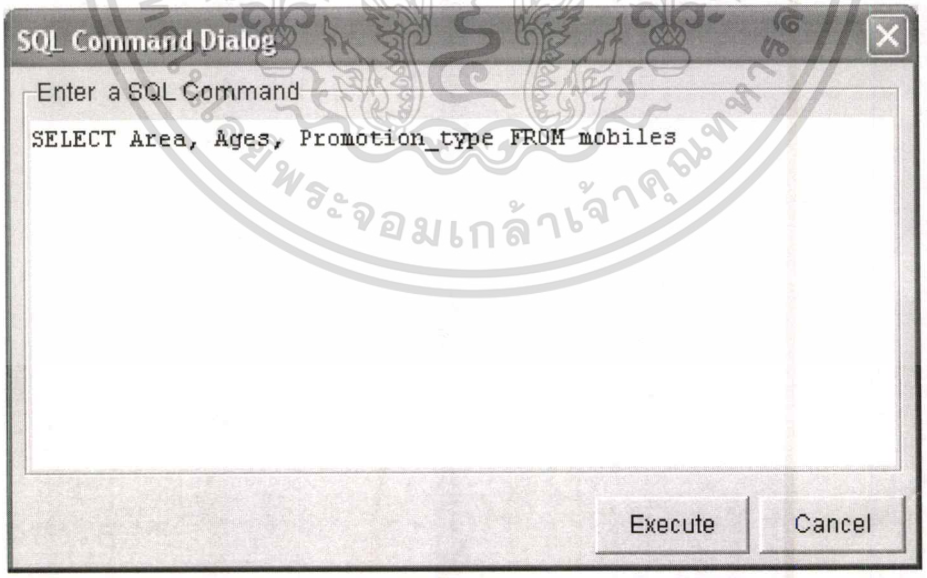
รูปที่ 5.1 หน้าจอหลัก และเมนูการทำงานของระบบ

- หน้าจอที่สองเป็นหน้าจอในส่วนของการติดต่อกับฐานข้อมูล รูปที่ 5.2 แสดงหน้าจอการติดต่อกับฐานข้อมูล โดยจะมี Drop Down List ให้เลือก Driver และ URL ของฐานข้อมูลที่ต้องการติดต่อ และถ้าฐานข้อมูลนั้นได้กำหนด User Name และ Password ไว้ ผู้ใช้จะต้องกรอก User Name และ Password ลงไปด้วย โดยฐานข้อมูลที่ระบบมีให้เลือกนั้นได้แก่ Microsoft Access และ JDataStore ซึ่งเป็นฐานข้อมูลของเครื่องมือพัฒนาระบบ JBuilder และเมื่อเสร็จสิ้นขั้นตอนดังกล่าวแล้วให้กดปุ่ม Connect เพื่อให้ระบบทำการติดต่อกับฐานข้อมูล

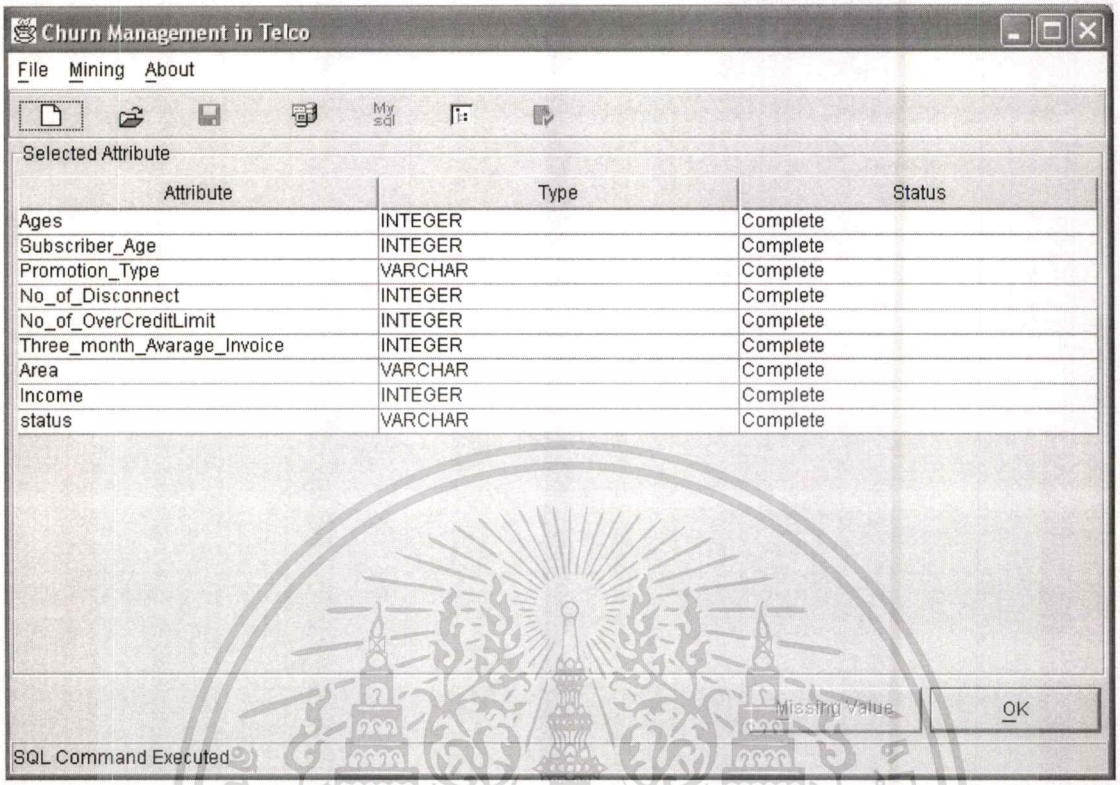


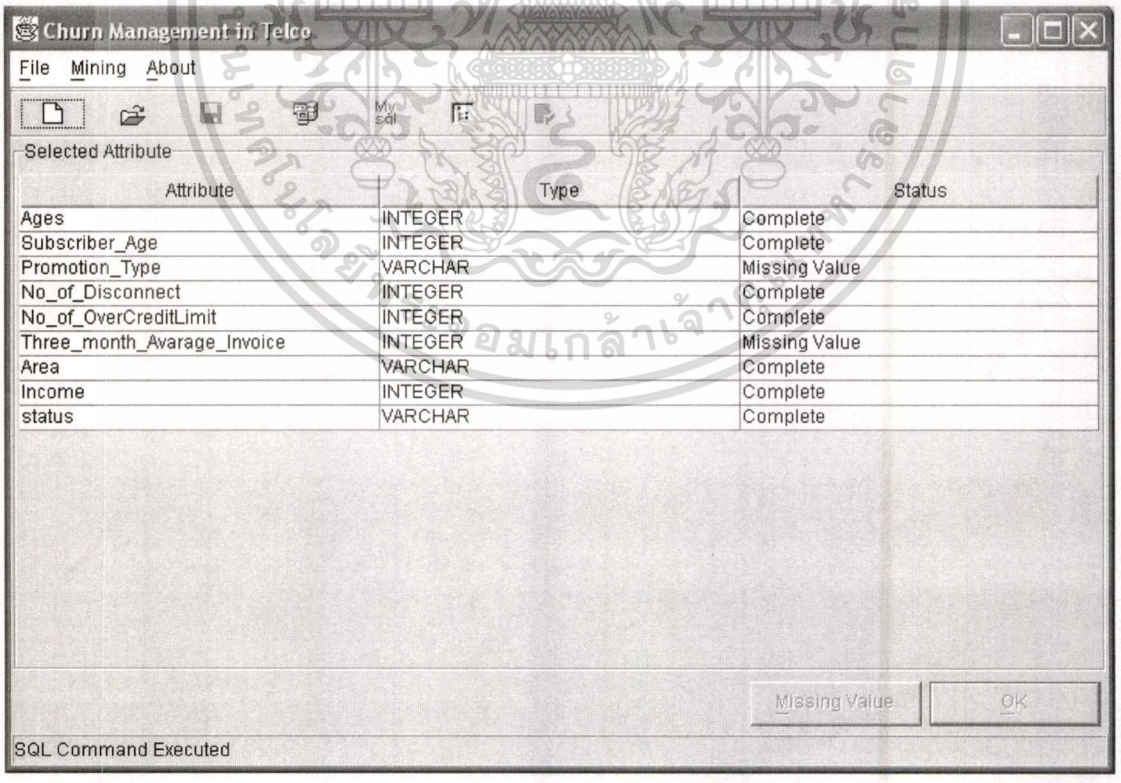
รูปที่ 5.2 หน้าจอการติดต่อกับฐานข้อมูล

- หน้าจอที่สามเป็นหน้าจอการเลือกข้อมูลจากฐานข้อมูลที่ติดต่อไว้แล้ว ด้วยคำสั่ง SQL โดยใช้คำสั่ง SQL ในการระบุแอตทริบิวต์ และตารางที่ต้องการ หน้าจอการเลือกข้อมูลด้วยคำสั่ง SQL แสดงในรูปที่ 5.3



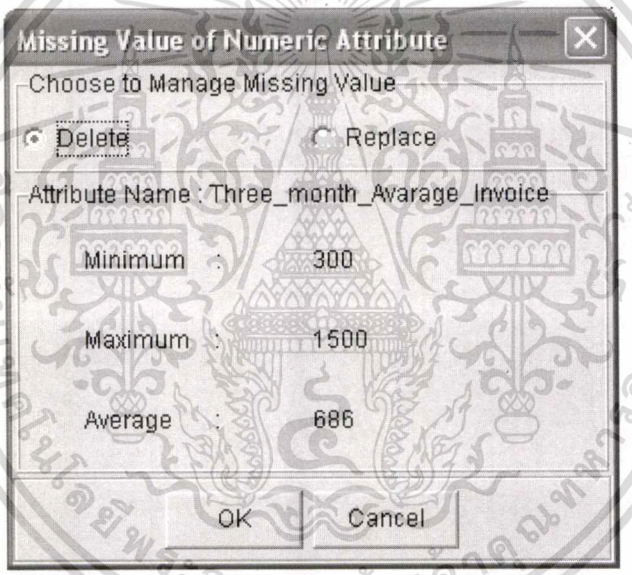
รูปที่ 5.3 หน้าจอการเลือกข้อมูลโดยใช้คำสั่ง SQL





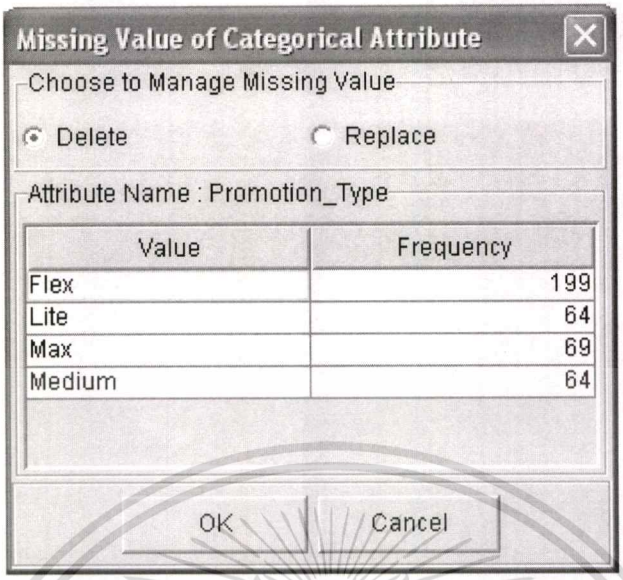
รูปที่ 5.4 หน้าจอแสดงตารางแอตทริบิวต์ที่เลือก และแสดงสถานะความสมบูรณ์ของข้อมูลที่เลือก

- หน้าจอที่สี่เป็นหน้าจอแสดงแอตทริบิวต์ที่เลือกไว้ โดยจะแสดงชื่อ (Name), ชนิด (Type) และสถานะ (Status) ของแอตทริบิวต์ โดยสถานะจะเป็นส่วนที่รายงานว่าข้อมูลในแต่ละแอตทริบิวต์นั้นสมบูรณ์หรือไม่กล่าวคือ ค่าในแอตทริบิวต์ขาดหายไป (Missing Value) หรือไม่ โดยถ้าทุกเรคคอร์ดของแอตทริบิวต์มีค่าครบจะแสดงสถานะ “Complete” ดังแสดงในรูปที่ 5.4 รูปบน แต่ถ้าแอตทริบิวต์มีค่าบางค่าขาดหายไป จะแสดงสถานะเป็น “Missing Value” ดังแสดงในรูปที่ 5.4 รูปล่าง และถ้ามีบางแอตทริบิวต์ที่มีค่าขาดหายไป ต้องทำการแก้ไข โดยเลือกปุ่ม “Missing Value” ซึ่งระบบจะมีหน้าจอแนะนำค่าที่จะใส่แทนที่ค่าที่ขาดหายไป ถ้าแอตทริบิวต์ที่มีค่าขาดหายไปมีชนิดเป็นตัวเลข(Numerical) ระบบจะหาค่าเฉลี่ยจากค่าสูงสุด และค่าต่ำสุดของแอตทริบิวต์นั้นๆ และหารด้วยจำนวนเรคคอร์ดที่ขาดหายไป ดังแสดงในรูปที่ 5.5

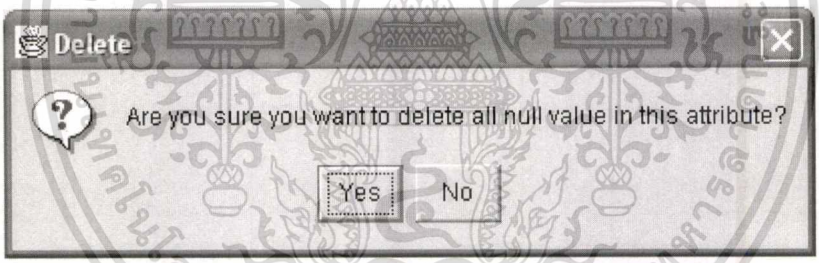


รูปที่ 5.5 หน้าจอแนะนำค่าที่จะใส่แทนค่าที่หายไปสำหรับแอตทริบิวต์แบบตัวเลข

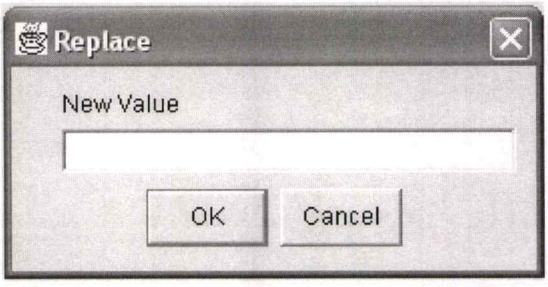
และถ้าแอตทริบิวต์ที่มีค่าขาดหายไปเป็นชนิดตัวอักษร(Categorical) ระบบจะแสดงค่าในแอตทริบิวต์ และความถี่ (Frequency) ของค่านั้นๆ ดังรูปที่ 5.6 โดยเลือก “Delete” และกดปุ่ม OK เพื่อทำการลบเรคคอร์ดที่มีค่าขาดหายไปทิ้ง ก่อนทำการลบเรคคอร์ดทิ้ง จะมีไดอะล็อกบ็อกซ์ถามยืนยันการลบอีกครั้งดังแสดงในรูปที่ 5.7 หรือเลือก “Replace” และกดปุ่ม OK เพื่อแทนที่ค่าที่ขาดหายไป ในรูปที่ 5.8 แสดงหน้าต่างให้ใส่ค่าใหม่ และกด OK เพื่อแทนที่ค่าใหม่ลงไป หลังจากแก้ไขข้อมูลที่ขาดหายไปครบทุกแอตทริบิวต์แล้ว กดปุ่ม OK เพื่อไปยังหน้าจอต่อไป



รูปที่ 5.6 หน้าจอในการจัดการค่าที่หายไปสำหรับแอตทริบิวต์แบบตัวอักษร

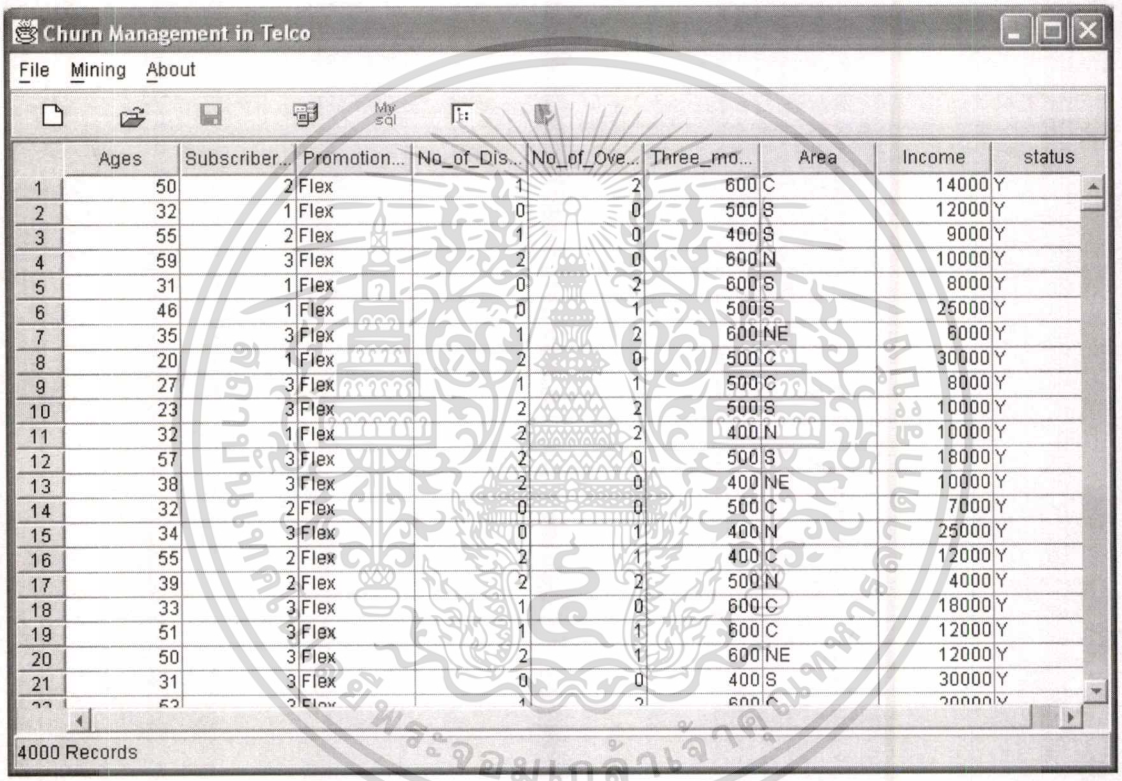


รูปที่ 5.7 ไดอะล็อกบ๊อกซ์ยืนยันการลบเรคคอร์ดที่ขาดหายไป



รูปที่ 5.8 ไดอะล็อกบ๊อกซ์กรอกค่าใหม่เพื่อแทนที่ค่าที่ขาดหายไป

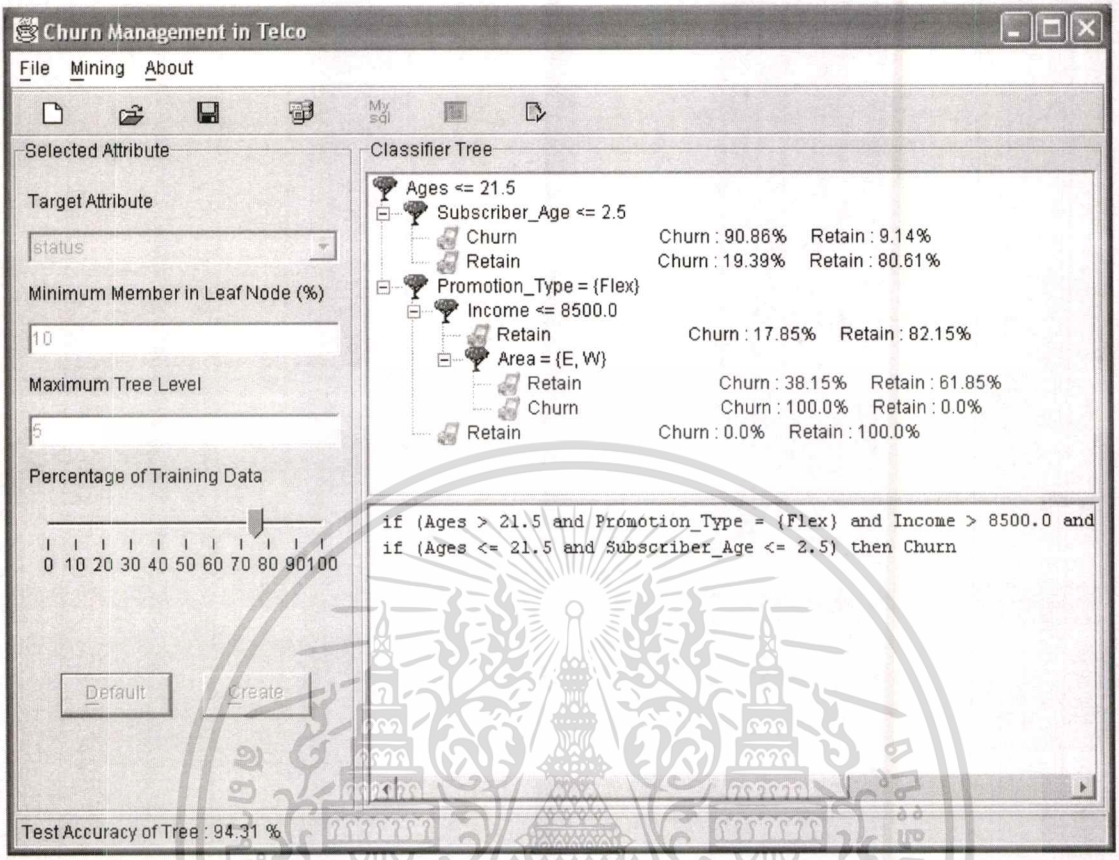
- หน้าจอที่ห้าเป็นหน้าจอแสดงตารางของข้อมูลที่เลือกมา โดยเป็นข้อมูลที่สมบูรณ์ ไม่มีส่วนของค่าที่ขาดหายไปอยู่ (Missing Value) ดังรูปที่ 5.9 และกดปุ่ม Mining หรือเลือกเมนู Mining เพื่อไปยังหน้าจอการสร้างโมเดลต่อไป
- หน้าจอที่หกในรูปที่ 5.10 เป็นหน้าจอของการสร้างโมเดล โดยก่อนทำการสร้างโมเดลต้องทำการกำหนดค่าต่างๆ ดังที่กล่าวไปแล้วในหัวข้อที่ 5.3.3 และกดปุ่ม “Create” เพื่อให้ระบบทำการสร้างโมเดล และแสดงผลในรูปโมเดลดิซิชั่นทรี และในรูปกฎการแบ่งกลุ่ม



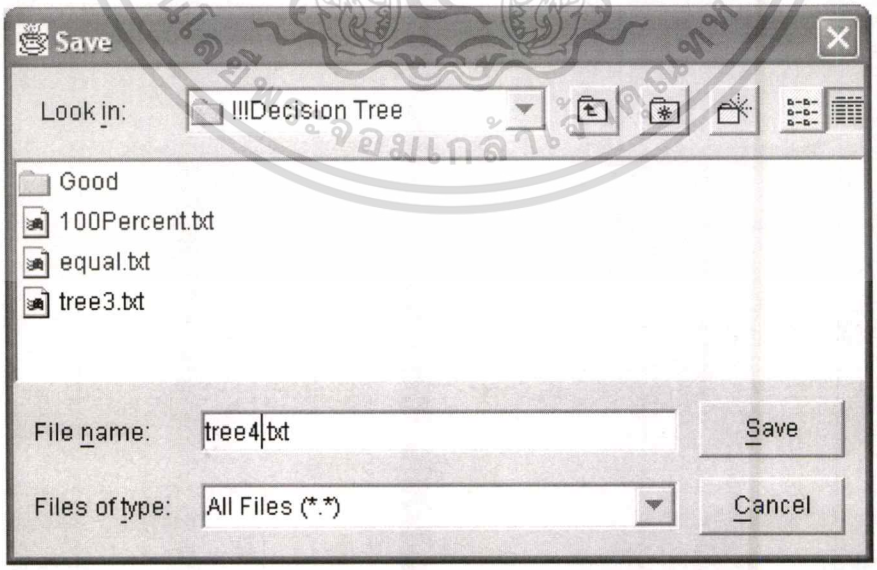
	Ages	Subscriber...	Promotion...	No_of_Dis...	No_of_Ove...	Three_mo...	Area	Income	status
1	50	2	Flex	1	2	600	C	14000	Y
2	32	1	Flex	0	0	500	S	12000	Y
3	55	2	Flex	1	0	400	S	9000	Y
4	59	3	Flex	2	0	600	N	10000	Y
5	31	1	Flex	0	2	600	S	8000	Y
6	46	1	Flex	0	1	500	S	25000	Y
7	35	3	Flex	1	2	600	NE	6000	Y
8	20	1	Flex	2	0	500	C	30000	Y
9	27	3	Flex	1	1	500	C	8000	Y
10	23	3	Flex	2	2	500	S	10000	Y
11	32	1	Flex	2	2	400	N	10000	Y
12	57	3	Flex	2	0	500	S	18000	Y
13	38	3	Flex	2	0	400	NE	10000	Y
14	32	2	Flex	0	0	500	C	7000	Y
15	34	3	Flex	0	1	400	N	25000	Y
16	55	2	Flex	2	1	400	C	12000	Y
17	39	2	Flex	2	2	500	N	4000	Y
18	33	3	Flex	1	0	600	C	18000	Y
19	51	3	Flex	1	1	600	C	12000	Y
20	50	3	Flex	2	1	600	NE	12000	Y
21	31	3	Flex	0	0	400	S	30000	Y

รูปที่ 5.9 หน้าจอแสดงตารางของข้อมูลที่เลือก

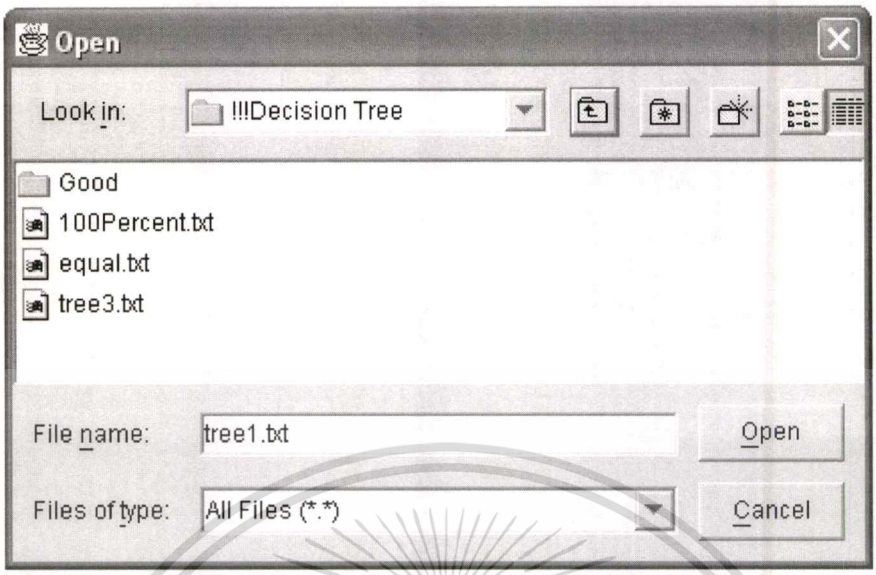
- หน้าจอที่เจ็ด ดังรูปที่ 5.11 เป็นหน้าจอในการเก็บบันทึกโมเดลที่สร้างเป็นไฟล์ โดยเก็บในรูปแบบ *.txt เพื่อนำโมเดลที่เคยสร้างไว้แล้วมาแสดงได้ใหม่โดยไม่ต้องสร้างโมเดลซ้ำใหม่อีกครั้ง โดยรูปที่ 5.12 และ 5.13 แสดงหน้าจอการเปิดไฟล์ของโมเดลที่เคยสร้างไว้ และแสดงโมเดลในแบบ Tree Graph และ Rule รวมทั้งแสดงรายละเอียดต่างๆ เกี่ยวกับโมเดล อาทิเช่น ค่าความถูกต้อง (Accuracy), ฐานข้อมูล ตาราง และแอตทริบิวต์ ที่เลือกมาใช้สร้างโมเดล, และวันที่สร้างโมเดลนี้



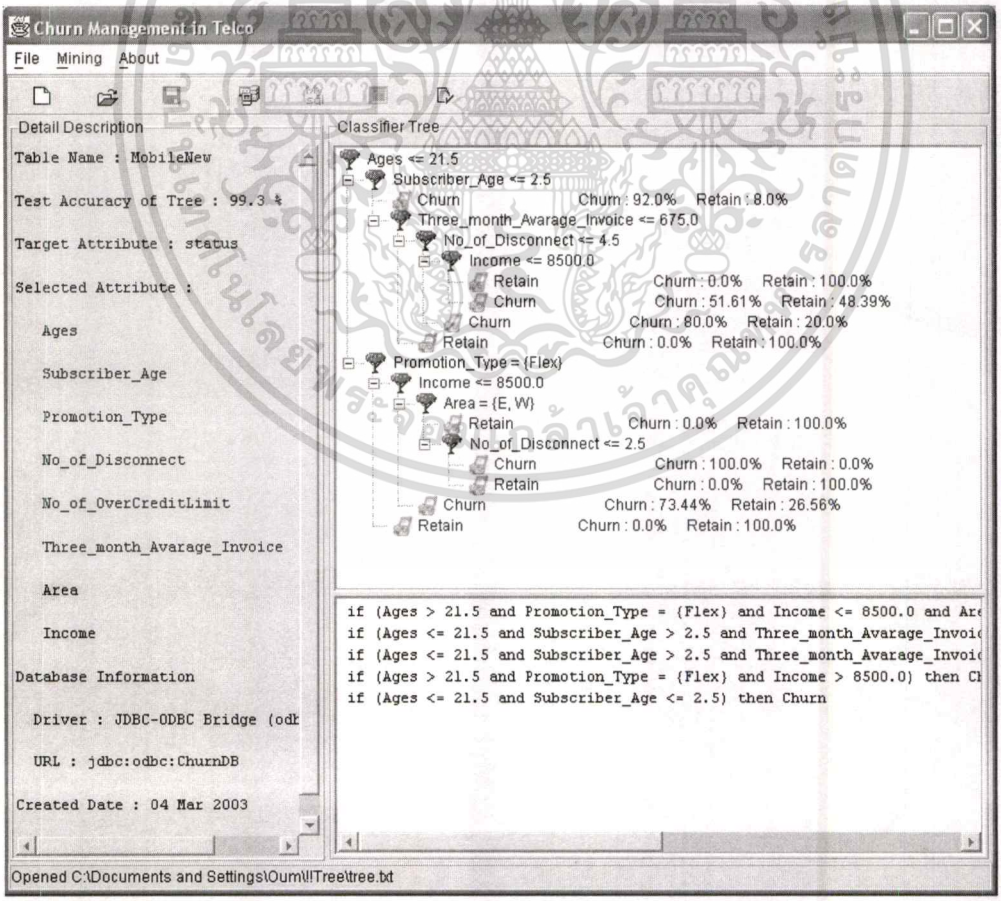
รูปที่ 5.10 หน้าจอการกำหนดค่าในการสร้างโมเดล และแสดงผลลัพธ์



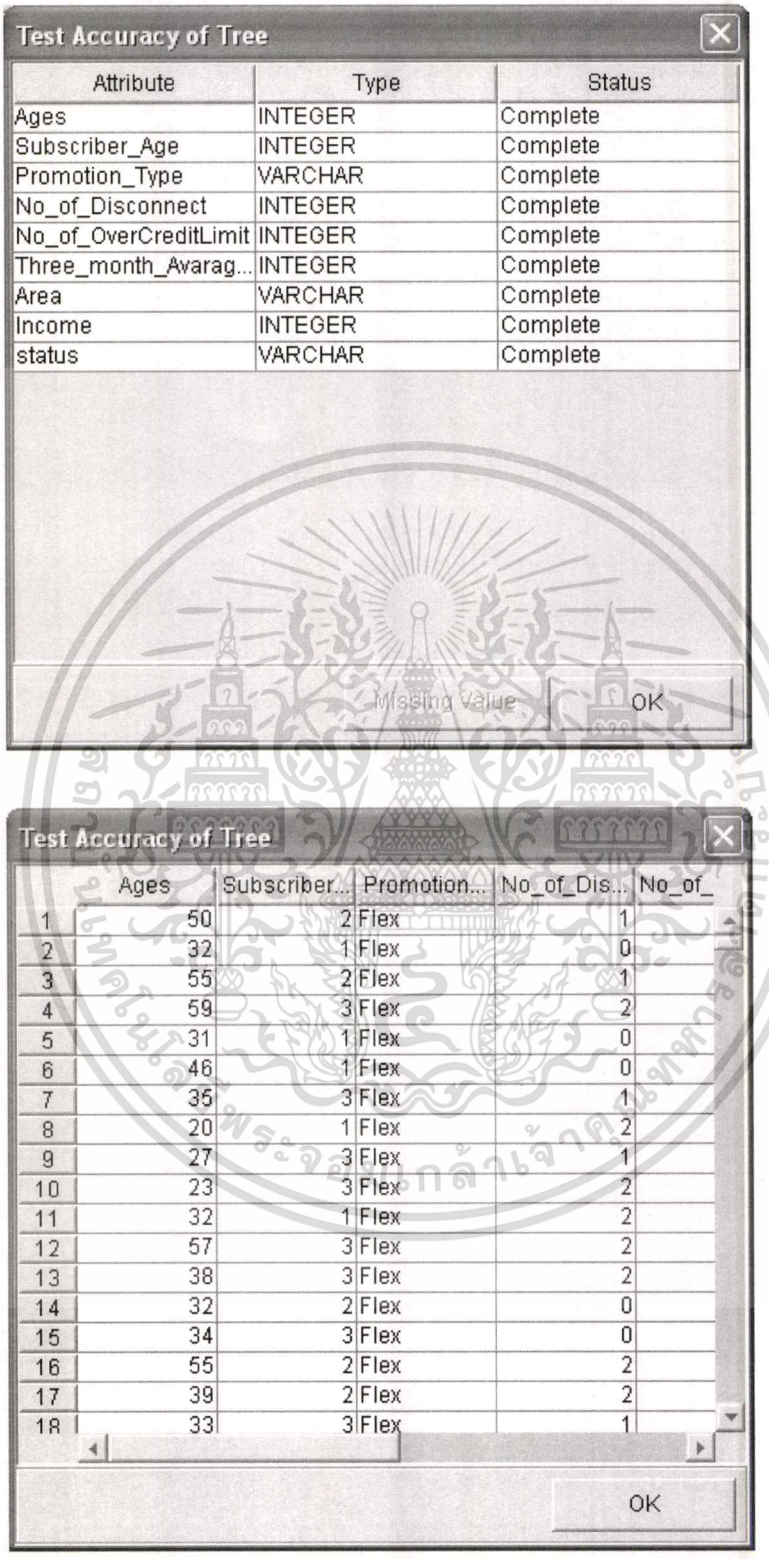
รูปที่ 5.11 หน้าจอการเก็บบันทึก (Save) โมเดล



รูปที่ 5.12 หน้าจอการเปิดไฟล์ (Open) โมเดลที่เคยสร้าง และบันทึกไว้

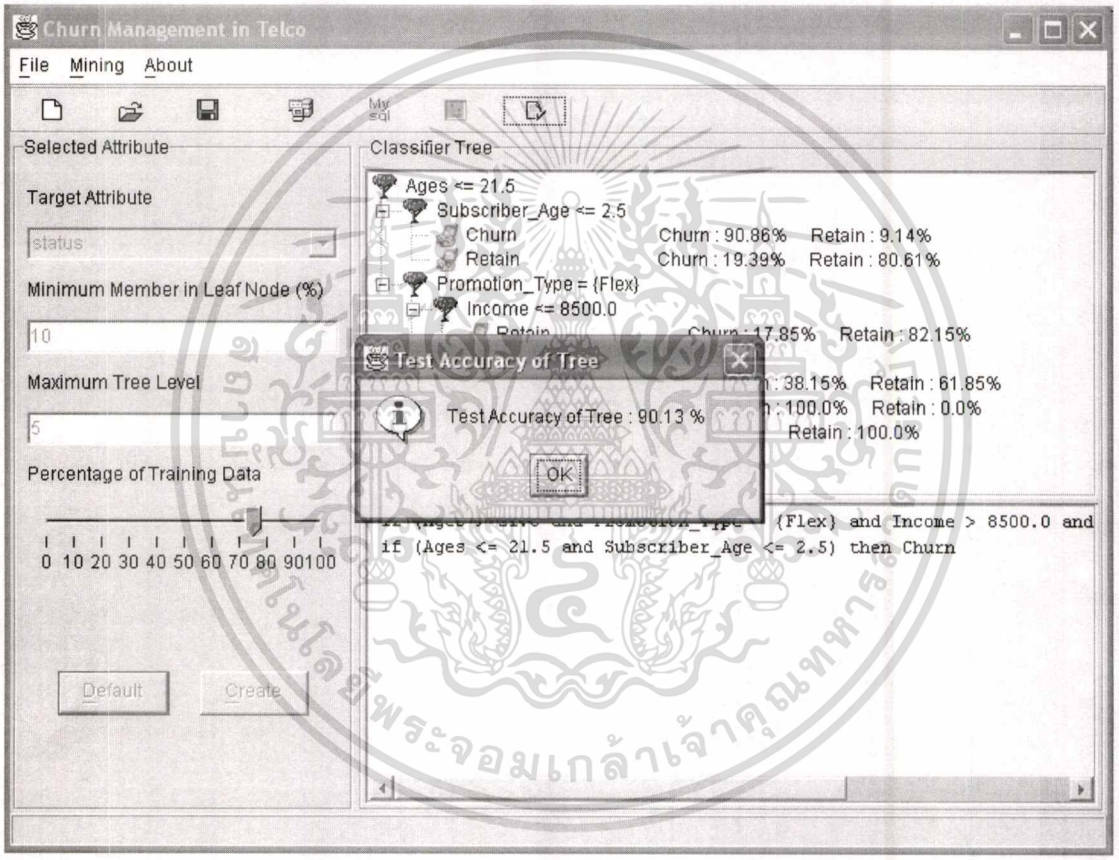


รูปที่ 5.13 หน้าจอแสดงโมเดลจากการเปิดไฟล์ และรายละเอียดต่างๆ เกี่ยวกับโมเดล



รูปที่ 5.14 หน้าจอแสดงตารางแอตทริบิวต์ และข้อมูล เพื่อใช้ในการทดสอบโมเดล

นอกจากนี้ โมเดลที่สร้างขึ้นมาแล้ว สามารถทดสอบค่าความถูกต้องโดยใช้ข้อมูลคนละชุด ซึ่งผู้ใช้จะทำการติดต่อกับฐานข้อมูลที่เก็บข้อมูลที่จะใช้ทดสอบ และเลือกข้อมูล โดยมีหน้าจอเหมือนกันกับการเลือกข้อมูลเพื่อสร้างโมเดล ดังแสดงในรูปที่ 5.14 หลังจากเลือกข้อมูล และเป็นข้อมูลที่สมบูรณ์แล้ว กดปุ่ม OK เพื่อให้ระบบนำข้อมูลดังกล่าวไปทำการทดสอบโมเดลที่ต้องการทดสอบ หลังจากนั้นระบบจะแสดงหน้าจอไดอะล็อกบ๊อกซ์ของค่าความถูกต้อง (Accuracy) ของโมเดลนั้นๆ ดังแสดงในรูปที่ 5.15



รูปที่ 5.15 ไดอะล็อกบ๊อกซ์แสดงค่าความถูกต้องที่ได้จากการทดสอบ

5.5 สรุปผลการทำงานของระบบ

ระบบนี้สามารถรองรับข้อมูลจำนวนมากๆ ได้โดยเวลาที่ใช้ในการประมวลผลค่อนข้างเร็ว และผลของโมเดลที่ได้จากระบบ จะแสดงค่าความถูกต้อง (Accuracy) อยู่ในรูปเปอร์เซ็นต์ เพื่อบอกความน่าเชื่อถือในการทำนายของโมเดล ซึ่งค่าความถูกต้องของโมเดลนั้นขึ้นอยู่กับปัจจัยต่างๆ

มากมาย หากเลือก และกำหนดค่าต่างๆ อย่างเหมาะสมแล้ว ก็จะได้ต้นไม้ที่มีค่าความถูกต้องที่ยอมรับได้ ปัจจัยต่างๆ เช่น

- จำนวนของข้อมูลที่เลือกมา ควรมีค่าที่จะใช้ในการทำนายว่าลูกค้ายกเลิกบริการ (Churn) หรือ ยังคงใช้บริการต่อไป (Retain) ในจำนวนที่ใกล้เคียงกัน
- การเลือกสัดส่วนของข้อมูลที่ใช้ในการสร้างต้นไม้ (Training Set) และส่วนของข้อมูลที่ใช้ทดสอบต้นไม้ (Test Set) ถ้ากำหนดสัดส่วนของ Training Set น้อยเกินไป อาจทำให้ค่าความถูกต้องของต้นไม้ลดน้อยลง เนื่องจากในการสร้างต้นไม้ ต้องทำการแบ่งข้อมูลจากส่วนของ Training Set ไปใช้ในการตัดกิ่ง (Pruning Set) ด้วย จึงทำให้ข้อมูลที่จะมาสร้างต้นไม้ลดน้อยลง ส่งผลให้ได้โมเดลที่ไม่มีความน่าเชื่อถือในการทำนายได้

บทที่ 6
สรุปผล

โครงการพัฒนาระบบนี้ทำการพัฒนาระบบ เพื่อประโยชน์ในการสนับสนุนการวิเคราะห์หาสาเหตุการยกเลิกบริการโทรศัพท์เคลื่อนที่ โดยในบทนี้เป็นการสรุปผลการศึกษา และข้อเสนอแนะในการพัฒนาระบบให้มีประสิทธิภาพ และตอบสนองต่อความต้องการในการใช้ระบบให้มากที่สุด

6.1 สรุปผลการศึกษา

ระบบการวิเคราะห์หาสาเหตุการยกเลิกบริการโทรศัพท์เคลื่อนที่ เป็นระบบที่อยู่ในรูปแบบของแอปพลิเคชั่นบนวินโดวส์ โดยระบบจะทำการสร้างโมเดล เริ่มจากการติดต่อกับฐานข้อมูล เพื่อเลือกข้อมูลที่ต้องการด้วยคำสั่ง SQL ซึ่งระบบสามารถรองรับข้อมูลได้ทั้งที่เป็นแอตทริบิวต์แบบ Numerical และ Categorical รวมทั้งสามารถจัดการกับค่าที่ขาดหายไป (Missing Value) และระบบจะนำข้อมูลที่เลือกไว้มาทำการสร้างโมเดล

ระบบสร้างโมเดลโดยใช้อัลกอริทึม SLIQ ซึ่งเป็นเทคนิคการแบ่งกลุ่ม (Classification) แบบ ดิซิชั่นทรีเพื่อทำการแบ่งกลุ่มข้อมูลตามกลุ่มที่ได้แบ่งไว้แล้ว คือกลุ่มลูกค้าที่ยกเลิกบริการ (Churn) และกลุ่มลูกค้าที่ยังคงใช้บริการต่อไป (Retain) ซึ่ง SLIQ สามารถขยายความสามารถในการรองรับกลุ่มของข้อมูลที่มีขนาดใหญ่ได้ โดย SLIQ ใช้เทคนิค Pre-sorting ในการสร้างต้นไม้ เพื่อลดเวลาในการประมวลผลกับข้อมูลที่เป็นตัวเลข ทำให้ช่วยลดเวลาประมวลผลโดยรวมของระบบ

ผลที่ได้จากระบบเป็นโมเดลการทำนายที่เป็น Decision Tree ซึ่งเป็นโมเดลที่สามารถเข้าใจได้ง่าย รวมทั้งยังแปลงไปเป็น SQL Statement ได้ง่ายอีกด้วย และระบบสามารถทำการบันทึกผลเป็นเท็กซ์ไฟล์เพื่อนำกลับมาเปิดดูโมเดลที่สร้างไว้ได้

6.2 ข้อเสนอแนะ

ดังที่กล่าวมาแล้ว โครงการพัฒนาระบบนี้ทำการพัฒนาระบบเพื่อประโยชน์ในการสนับสนุนการวิเคราะห์หาสาเหตุการยกเลิกบริการโทรศัพท์เคลื่อนที่ ดังนั้นระบบนี้สามารถนำไปพัฒนาให้มีประสิทธิภาพ และตรงกับความต้องการมากขึ้น โดยผู้พัฒนามีข้อเสนอแนะในการพัฒนาระบบดังนี้

- อาจมีการพัฒนาให้ระบบสามารถเก็บบันทึกผล โดยมีระบบป้องกันการแก้ไขข้อมูลในไฟล์เพื่อป้องกันการแก้ไขที่อาจเกิดขึ้นได้จากที่ได้กล่าวแล้วข้างต้น

- พัฒนาให้ระบบสามารถแปลงผลการทำนายให้อยู่ในรูปคำสั่ง SQL ซึ่งสามารถนำไปใช้ดึงข้อมูลจากฐานข้อมูลได้เลย โดยผู้ใช้ไม่ต้องมาแปลงเป็นคำสั่ง SQL เอง
- อาจมีการพัฒนาให้ระบบสามารถรวมเทคนิคดาต้าไมนิ่ง ซึ่งแต่ละเทคนิคมีข้อดีและข้อด้อยของแต่ละเทคนิคเอง การนำเทคนิคหลายๆ เทคนิคมารวมกันทำให้สามารถแก้ไขข้อด้อยของการที่ระบบมีเพียงเทคนิคเดียวได้
- ทำการพัฒนาระบบเป็นเว็บแอปพลิเคชั่น เพื่อความสะดวกและความยืดหยุ่นในการใช้งานมากยิ่งขึ้น

นอกเหนือจากการพัฒนาระบบให้มีประสิทธิภาพมากขึ้นแล้ว ข้อเสนอแนะอีกประการหนึ่งคือ ระบบนี้สามารถนำไปศึกษา เพื่อเป็นแนวทางในการพัฒนาการทำดาต้าไมนิ่งมาสร้างโมเดลการทำนายการยกเลิกบริการของลูกค้าในธุรกิจโทรศัพท์เคลื่อนที่ ให้มีประสิทธิภาพมากยิ่งขึ้น

ประวัติผู้เขียน

ชื่อ นามสกุล นางสาว อุรชา โพธิ์นิ่มแดง

วัน เดือน ปีเกิด 16 กุมภาพันธ์ พ.ศ. 2518

สถานที่เกิด จังหวัดกรุงเทพมหานคร

วุฒิการศึกษา ปริญญาตรีวิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยกรุงเทพ

บรรณานุกรม

J.C. Shafer, R. Agrawal, and M.Mehta. 1996. **"SPRINT: A Scalable Parallel Classifier for Data Mining"**. In Proc. of the 22nd VLDB Conference.

Jiawei, Han and Micheline, Kamber. 2000. **Data Mining Concepts and Techniques.** San Francisco: Morgan Kaufman.

M. Mehta, R. Agrwal, and J. Rissanen. 1996. **"SLIQ: A Fast Scalable Classifier for Data Mining"**. 18-32 In Proc. Of the fifth Int'l Conference on Extending Database Technology.

M. Mozer, R. Wolniewicz, D. Grimes, E. Johnson, H. Kaushansky. 2000. **"Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry"**. In IEEE Transactions on Neural Network.

Peter, Cabena. Et al. 1997. **Discovering Data Mining from Concept to Implementation.** New Jersey: Prentice Hall PTR.

Tom Soukup, Ian Davidson. 2002. **Visual Data Mining.** Canada: John Wiley & Sons, Inc.