**รายงานการวิเคราะห์ข้อมูลเบื้องต้น (Exploratory Data Analysis)**

**Telco Customer Churn Dataset**

**1. สรุปข้อมูลเบื้องต้น**

ชุดข้อมูล Telco Customer Churn ประกอบด้วย 7,043 เรคคอร์ด (rows) และ 21 คอลัมน์ (columns) โดยสามารถแบ่งประเภทตัวแปรได้ดังนี้:

* **ตัวแปรเชิงปริมาณ (Quantitative Variables) - 4 ตัวแปร:**
  + SeniorCitizen: สถานะผู้สูงอายุ (0/1)
  + tenure: ระยะเวลาการใช้บริการ (เดือน)
  + MonthlyCharges: ค่าบริการรายเดือน (ดอลลาร์)
  + TotalCharges: ค่าบริการรวมทั้งหมด (ดอลลาร์)
* **ตัวแปรเชิงคุณภาพ (Categorical Variables) - 17 ตัวแปร:**
  + ข้อมูลส่วนบุคคล: gender, Partner, Dependents
  + บริการโทรศัพท์: PhoneService, MultipleLines
  + บริการอินเทอร์เน็ต: InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies
  + ข้อมูลสัญญาและการชำระเงิน: Contract, PaperlessBilling, PaymentMethod
  + **ตัวแปรตาม: Churn (Yes/No)**

**โครงสร้างของข้อมูล**

|  |  |  |
| --- | --- | --- |
| Columns | ประเภทข้อมูล | คำอธิบาย |
| customer ID | String | รหัสลูกค้า |
| gender | Categorical | เพศ (Male/Female) |
| SeniorCitizen | Numerical | สถานะผู้สูงอายุ (1 = ใช่, 0 = ไม่ใช่) |
| Partner | Categorical | มีคู่ครอง (Yes/No) |
| Dependents | Categorical | มีผู้อยู่ในความดูแล (Yes/No) |
| tenure | Numerical | ระยะเวลาการใช้บริการ (เดือน) |
| PhoneService | Categorical | ใช้บริการโทรศัพท์ (Yes/No) |
| MultipleLines | Categorical | มีหลายเบอร์ (Yes/No) |
| InternetService | Categorical | ประเภทบริการอินเทอร์เน็ต (DSL, Fiber optic, No) |
| OnlineSecurity | Categorical | บริการรักษาความปลอดภัยออนไลน์ (Yes/No) |
| OnlineBackup | Categorical | บริการสำรองข้อมูลออนไลน์ (Yes/No) |
| DeviceProtection | Categorical | บริการคุ้มครองอุปกรณ์ (Yes/No) |
| TechSupport | Categorical | บริการสนับสนุนทางเทคนิค (Yes/No) |
| StreamingTV | Categorical | บริการสตรีมมิ่งทีวี (Yes/No) |
| StreamingMovies | Categorical | บริการสตรีมมิ่งภาพยนตร์ (Yes/No) |
| Contract | Categorical | ประเภทสัญญา (Month-to-month, One year, Two year) |
| PaperlessBilling | Categorical | ใบแจ้งหนี้แบบกระดาษ (Yes/No) |
| PaymentMethod | Categorical | วิธีการชำระเงิน |
| MonthlyCharges | Numerical | ค่าบริการรายเดือน |
| TotalCharges | Numerical | ค่าบริการรวมทั้งหมด |
| Churn | Categorical | สถานะการยกเลิกบริการ (Yes/No) |

**สถิติเชิงพรรณนา**

|  |  |  |  |
| --- | --- | --- | --- |
| Statistic | SeniorCitizen | tenure | MonthlyCharges |
| Count | 7043 | 7043 | 7043 |
| Mean | 0.16 | 32.37 | 64.76 |
| Std | 0.37 | 24.56 | 30.09 |
| Min | 0 | 0 | 18.25 |
| 25% | 0 | 9 | 35.5 |
| 50% | 0 | 29 | 70.35 |
| 75% | 0 | 55 | 89.85 |
| Max | 1 | 72 | 118.75 |

**การกระจายของตัวแปรเป้าหมาย (Churn):**

* ไม่ยกเลิกบริการ (No): 5,174 ราย (73.5%)
* ยกเลิกบริการ (Yes): 1,869 ราย (26.5%)

**การกระจายของประเภทสัญญา (Contract):**

* Month-to-month: 3,875 ราย (55.0%)
* Two year: 1,695 ราย (24.1%)
* One year: 1,473 ราย (20.9%)

A comparison of a graph

AI-generated content may be incorrect.

**ฮิสโตแกรมของ MonthlyCharges:**

กราฟแสดงการกระจายแบบ **Bimodal** โดยมียอดที่ประมาณ $20 และ $80 บ่งชี้ว่ามีลูกค้า 2 กลุ่มหลัก: กลุ่มใช้บริการน้อย และกลุ่มใช้บริการเต็มรูปแบบ

A graph of a graph showing a number of charge

AI-generated content may be incorrect.

**Box Plot ของ tenure vs Churn:**

A diagram of a box plot

AI-generated content may be incorrect.

ลูกค้าที่ยกเลิกบริการมี tenure เฉลี่ย **18.0** เดือน ในขณะที่ลูกค้าที่ไม่ยกเลิกมี tenure เฉลี่ย **37.6** เดือน แสดงให้เห็นชัดเจนว่าความภักดีของลูกค้า (tenure) มีความสัมพันธ์อย่างมากกับการคงอยู่ของลูกค้า

**2. วัตถุประสงค์ของการวิเคราะห์**

วัตถุประสงค์หลักของการวิเคราะห์นี้คือ เพื่อศึกษาปัจจัยที่ส่งผลต่อการยกเลิกการใช้บริการของลูกค้าโทรคมนาคม (Customer Churn) และพัฒนาโมเดลทำนายเพื่อช่วยธุรกิจลดการสูญเสียลูกค้าและเพิ่มรายได้

การวิเคราะห์จะเน้นไปที่การตอบคำถามสำคัญดังนี้:

* ลูกค้าประเภทไหนที่มีแนวโน้มจะยกเลิกบริการมากที่สุด?
* ปัจจัยใด (เช่น ระยะเวลาการใช้งาน, ประเภทสัญญา, บริการเสริม) ที่มีความสัมพันธ์กับการยกเลิกบริการมากที่สุด?
* เราสามารถสร้างโมเดลที่แม่นยำเพื่อทำนายลูกค้าที่มีความเสี่ยงสูงในการยกเลิกบริการล่วงหน้าได้หรือไม่?

**3. ประเภทของการเรียนรู้ (Supervised vs Unsupervised)**

ชุดข้อมูลนี้เหมาะสำหรับการเรียนรู้แบบมีผู้สอน (Supervised Learning) เนื่องจากมีตัวแปรเป้าหมาย Churn ที่บอกชัดเจนว่าลูกค้าแต่ละรายยกเลิกบริการ (Yes) หรือไม่ (No) นี่คือปัญหา Binary Classification ที่เราสามารถใช้ข้อมูลในอดีตของลูกค้าเพื่อฝึกโมเดลให้ทำนายพฤติกรรมการยกเลิกบริการของลูกค้าใหม่ได้

**4. อธิบายการทำงานของ fit method**

fit method คือขั้นตอนที่โมเดล Machine Learning เรียนรู้จากข้อมูลฝึก (training data) เหมือนกับการสอนเด็กให้รู้จักสิ่งต่างๆ

**อุปมา:** ลองนึกภาพว่าข้อมูล (data) คือ "คนในห้างสรรพสินค้า" และคุณสมบัติ (features) คือ "เสื้อผ้าที่พวกเขาสวมใส่" เมื่อเราให้โมเดลดูคนหลายพันคนพร้อมบอกว่าแต่ละคนชอบฟังเพลงประเภทไหน โมเดลจะค่อยๆ เรียนรู้ความสัมพันธ์จากรูปแบบที่ซ้ำๆ กันในข้อมูล เช่น "คนที่ใส่เสื้อสีดำ + กางเกงยีนส์ + รองเท้าบู๊ท มักชอบฟังเพลง Rock"

ในกรณีของลูกค้าโทรคมนาคม โมเดลจะเรียนรู้จากข้อมูล เช่น "ลูกค้าที่มี **tenure** สั้นๆ + จ่ายเงินเยอะ (MonthlyCharges สูง) + ไม่มีสัญญาผูกมัด (Contract = Month-to-month) มักจะยกเลิกบริการ (Churn = Yes)" กระบวนการเรียนรู้นี้จะปรับพารามิเตอร์ภายในโมเดลให้สามารถหาความสัมพันธ์และรูปแบบที่ซับซ้อนในข้อมูลได้

A diagram of a number of red and blue squares

AI-generated content may be incorrect.

**5. การตีความผลลัพธ์และเชื่อมโยงกับความเป็นจริง**

จากการวิเคราะห์ พบประเด็นสำคัญหลายจุดที่ควรให้ความสนใจ:

1. **อัตราการยกเลิกบริการสูง (26.5%):** นี่เป็นสัญญาณเตือนสำคัญ! ในทุกๆ 4 ลูกค้า จะมี 1 คนที่ยกเลิกบริการ ซึ่งหมายถึงการสูญเสียรายได้จำนวนมากและต้นทุนในการหาลูกค้าใหม่ที่สูง บริษัทควรใช้ผลการวิเคราะห์นี้ในการวางแผนแคมเปญรักษาลูกค้า (Retention Campaign)
2. **ลูกค้า Month-to-month มีความเสี่ยงสูง:** ลูกค้า 55% ไม่ผูกมัดสัญญาระยะยาว แสดงว่าพวกเขาสามารถเปลี่ยนไปใช้บริการคู่แข่งได้ง่าย บริษัทควรสร้างแรงจูงใจให้ลูกค้ากลุ่มนี้ทำสัญญายาวขึ้น เช่น เสนอส่วนลดหรือสิทธิพิเศษ
3. **ความสัมพันธ์ระหว่าง tenure และ Churn:** ลูกค้าใหม่ (tenure ต่ำ) มีแนวโน้มยกเลิกบริการมากกว่าอย่างชัดเจน (tenure เฉลี่ย 18.0 เดือน) หมายความว่าช่วง 6-12 เดือนแรกเป็นช่วงวิกฤตที่ต้องดูแลลูกค้าเป็นพิเศษ และทีมบริการลูกค้าควรให้ความสำคัญกับลูกค้าใหม่เป็นอันดับแรก
4. **การแยกกลุ่มลูกค้าชัดเจน:** จาก MonthlyCharges ที่มี 2 ยอด บ่งชี้ว่าบริษัทมีกลุ่มลูกค้า 2 แบบ: Basic User และ Premium User ควรมีกลยุทธ์การตลาดและข้อเสนอที่แตกต่างกันเพื่อตอบสนองความต้องการของแต่ละกลุ่ม

**6. อัลกอริทึม Machine Learning ที่ใช้**

ในการแก้ปัญหา Customer Churn ซึ่งเป็นปัญหาการจำแนกประเภท (Binary Classification) เราสามารถเลือกใช้อัลกอริทึมที่หลากหลายได้ ซึ่งแต่ละตัวมีจุดเด่นและหลักการทำงานที่เหมาะสมกับข้อมูลที่แตกต่างกันไป อัลกอริทึมทั้งหมดที่กล่าวถึงด้านล่างนี้อยู่ในกลุ่ม **Supervised Learning** เนื่องจากเราใช้ข้อมูลที่มีป้ายกำกับ (labeled data) อย่างตัวแปร Churn เพื่อสอนโมเดลให้เรียนรู้และทำนายผลลัพธ์ในอนาคต การเลือกใช้โมเดลหลายประเภทช่วยให้เราเปรียบเทียบประสิทธิภาพและเลือกโมเดลที่เหมาะสมที่สุดสำหรับชุดข้อมูลและวัตถุประสงค์ทางธุรกิจนี้ได้

**กลุ่มที่ทรงพลังและมักให้ผลลัพธ์ที่ดีที่สุดสำหรับข้อมูลตาราง (Tabular Data):**

* **XGBoost:** เป็นอัลกอริทึมในกลุ่ม Gradient Boosting ที่มีประสิทธิภาพสูงที่สุดตัวหนึ่งสำหรับข้อมูลในรูปแบบตาราง มีความแม่นยำสูง จัดการข้อมูลไม่สมดุลได้ดี (โดยเฉพาะอย่างยิ่งเมื่อมีลูกค้าที่ยกเลิกบริการน้อยกว่าลูกค้าทั่วไป) และประมวลผลได้รวดเร็ว เหมาะอย่างยิ่งสำหรับการค้นหาความสัมพันธ์ที่ซับซ้อนระหว่างตัวแปรต่างๆ เช่น ผลกระทบร่วมกันของ tenure ที่สั้นและ MonthlyCharges ที่สูง .
* **Random Forest:** เป็นอัลกอริทึมในกลุ่ม Ensemble Learning ที่ใช้การรวมกันของ Decision Tree หลายๆ ต้นเพื่อเพิ่มความแม่นยำและลดการ Overfitting สามารถบอกความสำคัญของตัวแปร (Feature Importance) ได้อย่างน่าเชื่อถือ ซึ่งเป็นประโยชน์อย่างมากในการระบุปัจจัยสำคัญที่สุดที่ส่งผลต่อการตัดสินใจของลูกค้า (เช่น ประเภทสัญญา หรือค่าบริการรายเดือน) เพื่อให้ธุรกิจสามารถวางแผนกลยุทธ์ได้อย่างตรงจุด .

**กลุ่มที่ใช้ได้ดีและมีคุณสมบัติเฉพาะตัว:**

* **SVM (Support Vector Machine):** เป็นอัลกอริทึมที่ทรงพลังในการจำแนกประเภท โดยเฉพาะอย่างยิ่งเมื่อข้อมูลมีมิติสูง (มีตัวแปรจำนวนมาก) ทำงานโดยการหาส่วนแบ่งที่เหมาะสมที่สุดเพื่อแยกกลุ่มลูกค้าออกจากกัน เหมาะสำหรับการสร้างโมเดลที่แม่นยำแม้ในกรณีที่ข้อมูลของลูกค้าที่ยกเลิกและไม่ยกเลิกบริการมีการทับซ้อนกัน
* **ANN (Artificial Neural Network):** เหมาะสำหรับปัญหาที่ต้องการความแม่นยำสูงและมีความซับซ้อน สามารถเรียนรู้ความสัมพันธ์ที่ไม่เป็นเชิงเส้น (non-linear relationships) ได้ดีกว่าโมเดลแบบดั้งเดิม ซึ่งอาจค้นพบรูปแบบที่ซ่อนอยู่ในข้อมูลที่อัลกอริทึมอื่นมองข้ามไป ตัวอย่างเช่น อาจเรียนรู้ว่าความสัมพันธ์ระหว่าง MonthlyCharges และ Churn นั้นแตกต่างกันไปตามกลุ่มลูกค้าที่ใช้บริการอินเทอร์เน็ตประเภทต่างๆ

**7. สรุปและข้อเสนอ**

การวิเคราะห์เบื้องต้นชี้ให้เห็นว่าชุดข้อมูลนี้มีศักยภาพสูงในการสร้างโมเดลทำนาย Customer Churn ที่มีประสิทธิภาพ ประเด็นสำคัญที่ควรดำเนินการต่อไป:

1. **การเตรียมข้อมูล (Data Preprocessing):** จัดการกับค่าว่าง (ถ้ามี), แปลงข้อมูลเชิงคุณภาพให้อยู่ในรูปแบบที่โมเดลสามารถเข้าใจได้ (เช่น One-Hot Encoding)
2. **การจัดการ Class Imbalance**: ใช้เทคนิคอย่าง SMOTE (Synthetic Minority Over-sampling Technique) เพื่อเพิ่มจำนวนข้อมูลของลูกค้าที่ Churn (Yes) ให้มีความสมดุลกับลูกค้าที่ไม่ Churn (No) เพื่อให้โมเดลเรียนรู้ทั้งสองคลาสอย่างสมบูรณ์
3. **การวิศวกรรมฟีเจอร์ (Feature Engineering):** สร้างตัวแปรใหม่ที่อาจมีประโยชน์ เช่น อัตราส่วน TotalCharges / tenure เพื่อหาค่าใช้จ่ายเฉลี่ยต่อเดือนของลูกค้า
4. **การประเมินผล (Model Evaluation):** ใช้เทคนิค Cross-Validation เพื่อประเมินประสิทธิภาพของโมเดลอย่างเชื่อถือได้ และใช้ Metric ที่เหมาะสมกับปัญหา Imbalanced Data เช่น F1-Score, Precision, Recall

สิ่งสำคัญที่สุดคือการนำผลการทำนายไปประยุกต์ใช้กับธุรกิจจริง (Business Application) เช่น การสร้างระบบแจ้งเตือนลูกค้าที่มีความเสี่ยงสูง เพื่อให้ทีมบริการลูกค้าสามารถเข้าหาและเสนอข้อเสนอพิเศษเพื่อรักษาลูกค้าไว้ได้ ซึ่งจะนำไปสู่การลดการสูญเสียลูกค้าและเพิ่มผลกำไรในระยะยาวได้อย่างยั่งยืน