**รายงานการวิเคราะห์ข้อมูลเบื้องต้น (Exploratory Data Analysis)**

**Telco Customer Churn Dataset**

**1. สรุปข้อมูลเบื้องต้น**

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

**การแยกประเภทตัวแปร:**

**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
* **ตัวแปรเป้าหมาย**: Churn (Yes/No)

โครงสร้างของข้อมูล

|  |  |  |
| --- | --- | --- |
| Columns | ประเภทข้อมูล | คำอธิบาย |
| customer ID |  |  |
| gender |  |  |
| SeniorCitizen |  |  |
| Partner |  |  |
| Dependents |  |  |
| tenure |  |  |
| PhoneService |  |  |
| MultipleLines |  |  |
| InternetService |  |  |
| OnlineSecurity |  |  |
| OnlineBackup |  |  |
| DeviceProtection |  |  |
| TechSupport |  |  |
| StreamingTV |  |  |
| StreamingMovies |  |  |
| Contract |  |  |
| PaperlessBilling |  |  |
| PaymentMethod |  |  |
| MonthlyCharges |  |  |
| TotalCharges |  |  |
| Churn |  |  |

สถิติเชิงพรรณนา

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

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

ชุดข้อมูลนี้เหมาะสำหรับ **การเรียนรู้แบบมีผู้สอน (Supervised Learning)** เนื่องจากมีตัวแปรเป้าหมาย Churn ที่บอกชัดเจนว่าลูกค้าแต่ละรายยกเลิกบริการ (Yes) หรือไม่ (No)

นี่เป็นปัญหา **Binary Classification** ที่เราสามารถใช้ข้อมูลในอดีตของลูกค้าเพื่อฝึกโมเดลให้ทำนายพฤติกรรมการยกเลิกบริการของลูกค้าใหม่ได้

**3. อธิบายการทำงานของ fit method**

**fit method** คือขั้นตอนที่โมเดล Machine Learning เรียนรู้จากข้อมูลฝึก เหมือนกับการสอนเด็กให้รู้จักสิ่งต่างๆ

**การอธิบายด้วยอุปมา:**

ลองนึกภาพว่า **ข้อมูล (data)** คือ **"คนในห้างสรรพสินค้า"** และ **คุณสมบัติ (features)** คือ **"เสื้อผ้าที่พวกเขาสวมใส่"**

เมื่อเราให้โมเดลดูคนหลายพันคนพร้อมบอกว่าแต่ละคนชอบฟังเพลงประเภทไหน โมเดลจะค่อยๆ เรียนรู้ว่า:

* "คนที่ใส่เสื้อสีดำ + กางเกงยีนส์ + รองเท้าบู๊ท มักชอบฟังเพลง Rock"
* "คนที่ใส่ชุดสูท + รองเท้าหนัง มักชอบฟังเพลง Jazz"

ในกรณีของลูกค้าโทรคมนาคม โมเดลจะเรียนรู้ว่า "ลูกค้าที่ใช้บริการสั้นๆ + จ่ายเงินเยอะ + ไม่มีสัญญาผูกมัด มักจะยกเลิกบริการ"

**4. ผลการวิเคราะห์เบื้องต้น (EDA Output)**

**ตัวอย่างข้อมูล 5 แถวแรก (head()):**

customerID gender SeniorCitizen Partner tenure MonthlyCharges TotalCharges Churn

7590-VHVEG Female 0 Yes 1 29.85 29.85 No

5575-GNVDE Male 0 No 34 56.95 1889.50 No

3668-QPYBK Male 0 No 2 53.85 108.15 Yes

7795-CFOCW Male 0 No 45 42.30 1840.75 No

9237-HQITU Female 0 No 2 70.70 151.65 Yes

**จำนวนตัวอย่าง (nrow()): 7,043 เรคคอร์ด**

**สรุปสถิติตัวแปรเชิงปริมาณ (summary()):**

| **ตัวแปร** | **ค่าต่ำสุด** | **Q1** | **ค่าเฉลี่ย** | **มัธยฐาน** | **Q3** | **ค่าสูงสุด** |
| --- | --- | --- | --- | --- | --- | --- |
| tenure | 0 | 9.0 | 32.37 | 29.0 | 55.0 | 72 |
| MonthlyCharges | 18.25 | 35.50 | 64.76 | 70.35 | 89.85 | 118.75 |
| TotalCharges | 18.80 | 401.45 | 2283.30 | 1397.48 | 3794.74 | 8684.80 |

**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%)

**ฮิสโตแกรมของ MonthlyCharges:**

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

**Box Plot ของ tenure vs Churn:**

ลูกค้าที่ยกเลิกบริการมี tenure เฉลี่ย 17.9 เดือน ในขณะที่ลูกค้าที่ไม่ยกเลิกมี tenure เฉลี่ย 37.6 เดือน แสดงว่าความภักดีมีความสัมพันธ์กับการคงอยู่

**5. วัตถุประสงค์ของการวิเคราะห์**

**"เพื่อศึกษาปัจจัยที่ส่งผลต่อการยกเลิกการใช้บริการของลูกค้าโทรคมนาคม (Customer Churn) และพัฒนาโมเดลทำนายเพื่อช่วยธุรกิจลดการสูญเสียลูกค้าและเพิ่มรายได้"**

**6. การตีความผลลัพธ์และเชื่อมโยงกับความเป็นจริง**

จากการวิเคราะห์ พบประเด็นสำคัญหลายจุด:

**1. อัตราการยกเลิกบริการสูง (26.5%)**  
นี่เป็นสัญญาณเตือนสำคัญ! ในทุกๆ 4 ลูกค้า จะมี 1 คนที่ยกเลิกบริการ บริษัทกำลังสูญเสียรายได้และต้นทุนการหาลูกค้าใหม่สูง

**2. ลูกค้า Month-to-month มีความเสี่ยงสูง**  
ลูกค้า 55% ไม่ผูกมัดสัญญาระยะยาว แสดงว่าพวกเขาสามารถเปลี่ยนไปใช้บริการคู่แข่งได้ง่าย บริษัทควรสร้างแรงจูงใจให้ลูกค้าทำสัญญายาวขึ้น

**3. ความสัมพันธ์ระหว่าง tenure และ Churn**  
ลูกค้าใหม่ (tenure ต่ำ) มีแนวโน้มยกเลิกบริการมากกว่า หมายความว่าช่วง 6-12 เดือนแรกเป็นช่วงวิกฤตที่ต้องดูแลลูกค้าเป็นพิเศษ

**4. การแยกกลุ่มลูกค้าชัดเจน**  
จาก MonthlyCharges ที่มี 2 ยอด บ่งชี้ว่าบริษัทมีกลุ่มลูกค้า 2 แบบ: Basic User และ Premium User ควรมีกลยุทธ์การตลาดแยกกัน

**7. อัลกอริทึม Machine Learning ที่แนะนำ**

**1. Decision Tree**

**เหตุผล:**

* ตีความผลได้ง่าย สามารถอธิบายให้ทีมธุรกิจฟังได้
* จัดการตัวแปรผสม (Numerical + Categorical) ได้ดี
* แสดง Decision Rules ที่ชัดเจน เช่น "ถ้า tenure < 12 เดือน AND Contract = Month-to-month แล้วมีโอกาส Churn สูง"

**2. Random Forest**

**เหตุผล:**

* ประสิทธิภาพสูงกว่า Decision Tree เพราะใช้หลายต้นไม้มาโหวต
* ให้ Feature Importance ที่เชื่อถือได้ บอกว่าตัวแปรไหนสำคัญที่สุด
* ทนต่อ Outliers และไม่ Overfit ง่าย
* เหมาะสำหรับข้อมูลที่มี Mixed Data Types เช่นนี้

**3. XGBoost**

**เหตุผล:**

* มีประสิทธิภาพสูงที่สุดสำหรับข้อมูล Tabular
* จัดการ Imbalanced Data (73:27) ได้ดีด้วยพารามิเตอร์ scale\_pos\_weight
* มี Early Stopping ป้องกัน Overfitting
* เหมาะสำหรับการใช้งานจริงที่ต้องการความแม่นยำสูง

**สรุปและข้อเสนอแนะ**

การวิเคราะห์เบื้องต้นชี้ให้เห็นว่าชุดข้อมูลนี้มีศักยภาพสูงในการสร้างโมเดลทำนาย Customer Churn ประเด็นสำคัญที่ควรดำเนินการต่อไป:

1. **ใช้ SMOTE** จัดการ Class Imbalance เพื่อให้โมเดลเรียนรู้ทั้งสองคลาสอย่างสมดุล
2. **Feature Engineering** สร้างตัวแปรใหม่ เช่น อัตราส่วน TotalCharges/tenure
3. **Cross Validation** ประเมินประสิทธิภาพอย่างเชื่อถือได้
4. **Business Application** นำผลการทำนายไปใช้วางแผน Retention Campaign

โดยสิ่งสำคัญที่สุดคือไม่ใช่แค่สร้างโมเดลที่แม่นยำ แต่ต้องสร้างโมเดลที่ธุรกิจนำไปใช้ได้จริงและเกิดประโยชน์ในการลดการสูญเสียลูกค้า