Data-Driven Growth

สถานการณ์:

น้องคือ **Data Analyst คนใหม่ของบริษัท "Alpha-Commerce"** ซึ่งเป็นบริษัท Startup ด้าน E-commerce ที่กำลังเติบโตอย่างรวดเร็ว บริษัทต้องการที่จะก้าวไปอีกขั้นด้วยการใช้ข้อมูลเพื่อตัดสินใจทางธุรกิจให้มีประสิทธิภาพมากขึ้น

ผู้บริหารได้มอบหมายโปรเจกต์สำคัญให้น้อง เพื่อวิเคราะห์ยอดขายและพฤติกรรมลูกค้าในช่วง 3 เดือนที่ผ่านมา (มิถุนายน-สิงหาคม 2567) โดยมีเป้าหมายหลักคือ **"การทำความเข้าใจลูกค้าและยอดขายให้ลึกซึ้ง และสร้างโมเดลทำนายโอกาสที่ลูกค้าจะกลับมาซื้อซ้ำ"**

ความท้าทาย: ข้อมูลที่น้องได้รับมานั้นเป็นข้อมูลดิบที่ถูกส่งออกมาจากระบบ ทำให้มีความยุ่งเหยิงและไม่สมบูรณ์อยู่หลายจุด น้องจะต้องใช้ทักษะทั้งหมดที่ได้เรียนมา ตั้งแต่การทำความสะอาดข้อมูล ไปจนถึงการวิเคราะห์เชิงลึกและการสร้าง Machine Learning Model เพื่อตอบคำถามสำคัญของผู้บริหาร

เป้าหมายของโปรเจกต์:

1. **การวิเคราะห์เชิงลึก:** นำเสนอข้อมูลเชิงลึกเกี่ยวกับยอดขายและพฤติกรรมลูกค้าผ่าน Visualizations ที่ชัดเจนและน่าสนใจ
2. **การจัดกลุ่มลูกค้า:** สร้างโมเดลการจัดกลุ่มลูกค้า (Clustering) เพื่อแบ่งลูกค้าออกเป็นกลุ่มตามพฤติกรรมการซื้อ
3. **การทำนาย:** สร้างโมเดลทำนายโอกาสที่ลูกค้าจะเป็นกลุ่ม **"High-Spender"** (ลูกค้าที่ใช้จ่ายสูง) เพื่อสนับสนุนการตัดสินใจของฝ่ายการตลาด
4. **ข้อเสนอแนะทางธุรกิจ:** สรุปผลการวิเคราะห์ทั้งหมดเป็นข้อเสนอแนะที่สามารถนำไปปฏิบัติได้จริง

โดยโปรเจกต์นี้จะแบ่งออกเป็น 2 ส่วนหลัก คือ **"การวิเคราะห์ข้อมูล"** และ **"การสร้างโมเดล Machine Learning"** ซึ่งน้องจะต้องส่งมอบทั้งในรูปแบบของ Presentation และโค้ดที่สามารถรันได้จริง

ไฟล์ csv ที่แนบมาด้วย: กดลิงก์ เพื่อเข้าดู

ไฟล์ที่ 1: order_details.csv

ไฟล์นี้เป็นข้อมูลรายละเอียดของคำสั่งซื้อทั้งหมดในช่วงเวลาหนึ่ง มีคอลัมน์และปัญหาที่อาจพบดังนี้:

- **order_id:** รหัสคำสั่งซื้อ (ไม่ซ้ำกัน)
- **order_date:** วันที่สั่งซื้อ (อาจมี format ที่ไม่สอดคล้องกัน เช่น YYYY-MM-DD และ DD/MM/YYYY)
- **Customer ID:** รหัสลูกค้า (อาจมีช่องว่างนำหน้า/ต่อท้าย หรือมีรูปแบบที่แตกต่างกัน เช่น C0001 และ ID_001)
- **product_category:** ประเภทสินค้า (อาจมีช่องว่างนำหน้า/ต่อท้าย หรือมีค่าที่ผิดปกติ เช่น N/A)
- **product_price:** ราคาต่อหน่วยของสินค้า (อาจมีค่าว่าง NaN หรือข้อมูลที่ไม่ถูกต้อง)
- **quantity:** จำนวนสินค้าที่สั่งซื้อ (อาจมีค่าที่ผิดปกติ เช่น 0)
- **total_price:** ราคารวมของสินค้าในคำสั่งซื้อ (คำนวณจาก product_price * quantity)

ไฟล์ที่ 2: customer_profiles.csv

ไฟล์นี้เป็นข้อมูลโปรไฟล์ของลูกค้าแต่ละคน มีคอลัมน์และปัญหาที่อาจพบดังนี้:

- **CUSTOMER_ID:** รหัสลูกค้า (อาจมีช่องว่างนำหน้า/ต่อท้าย และมีรูปแบบที่แตกต่างกัน)
- **city:** เมืองที่ลูกค้าอาศัยอยู่ (อาจมีช่องว่างนำหน้า/ต่อท้าย)
- **membership_level:** ระดับสมาชิก (Bronze, Silver, Gold)
- **registration_date:** วันที่ลงทะเบียนเป็นสมาชิก (อาจมีค่าที่ผิดปกติ เช่น unknown)

ไฟล์ที่ 3: ml_ready_features.csv

ไฟล์นี้เป็นผลลัพธ์จากการทำ Feature Engineering ซึ่งเป็นการนำข้อมูลจาก 2 ไฟล์ข้างต้นมารวมและสร้างเป็นคุณสมบัติใหม่สำหรับลูกค้าแต่ละคนแล้ว ไฟล์นี้จะถูกนำไปใช้ใน ชุดคำถามที่ 2: การสร้าง Machine Learning Model โดยตรง

คอลัมน์สำคัญในไฟล์นี้ได้แก่:

- **customer_id: รหัสลูกค้า**
- **total_spend_last_3_months:** ยอดใช้จ่ายรวมของลูกค้าในช่วง 3 เดือนที่ผ่านมา
- **purchase_count_last_3_months:** จำนวนครั้งที่ลูกค้าซื้อในช่วง 3 เดือนที่ผ่านมา
- **avg_spend_per_purchase:** ยอดใช้จ่ายเฉลี่ยต่อคำสั่งซื้อ
- **customer_lifetime:** อายุการเป็นลูกค้าในหน่วยวัน
- **membership_level:** ระดับสมาชิก
- **most_purchased_category:** ประเภทสินค้าที่ลูกค้าซื้อบ่อยที่สุด
- **repeat_purchase:** การที่ลูกค้ากลับมาซื้อสินค้าอะไรก็ได้ในเดือนถัดไป (1: ถ้าลูกค้ากลับมาซื้อสินค้าอะไรก็ได้ในเดือนกันยายน ซึ่งในที่นี้เราไม่ได้สนใจ)

แนวทางการใช้ไฟล์ในการทำโปรเจกต์

- ขั้นตอนการทำความสะอาดข้อมูล (Data Cleaning): น้องต้องจัดการปัญหาด้านคุณภาพของข้อมูลในไฟล์ **order_details.csv** และ **customer_profiles.csv** เพื่อใช้ในการวิเคราะห์ในชุดที่ 1
- การวิเคราะห์ในชุดที่ 1: ใช้ข้อมูลที่ผ่านการทำความสะอาดและรวมกันแล้วจาก 2 ไฟล์ข้างต้นเพื่อตอบคำถามเชิงวิเคราะห์และสร้าง Visualization ต่างๆ
- การสร้างโมเดลในชุดที่ 2: น้องจะเริ่มทำโจทย์ในส่วนนี้โดยการโหลดไฟล์ **ml_ready_features.csv** เข้าสู่โปรแกรมโดยตรง และใช้คอลัมน์ต่างๆ ที่สร้างไว้แล้วเพื่อสร้างโมเดลทั้ง Logistic Regression และ K-Means Clustering

---

ชุดคำถามที่ 1: การ Data Analysis และ Visualization

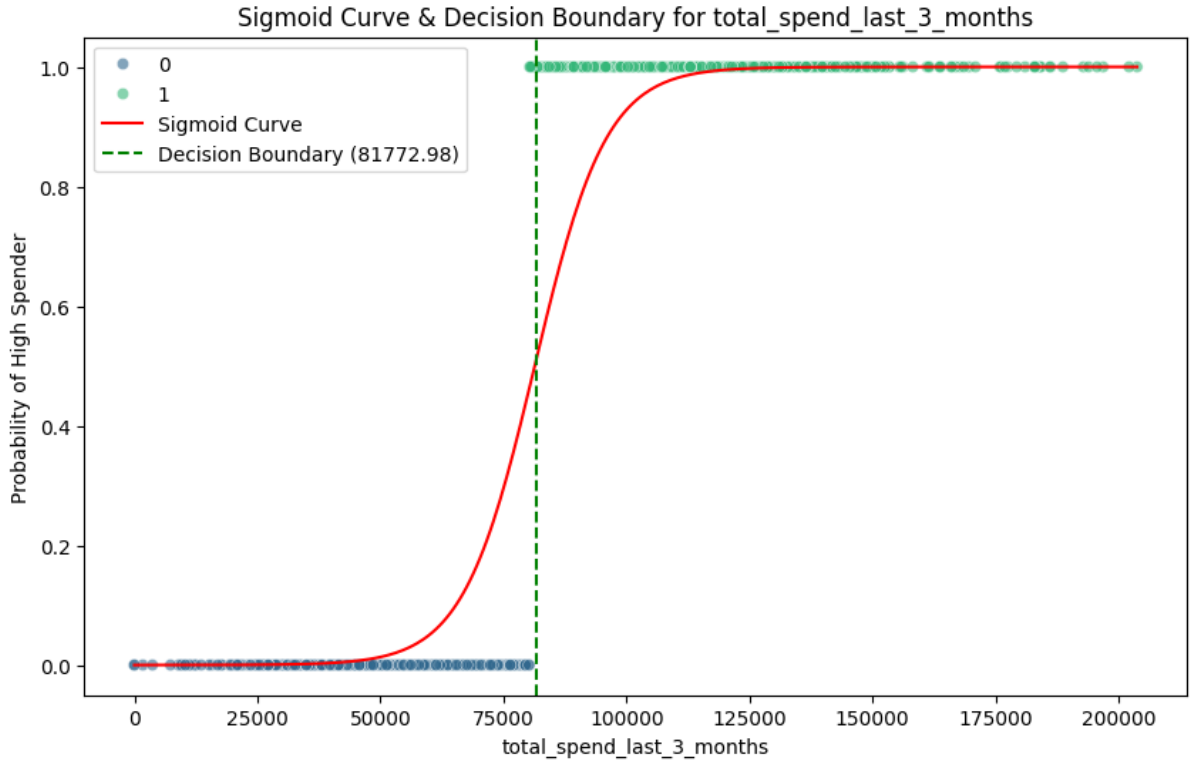
1. นำเข้าไฟล์ order_details.csv และ customer_profiles.csv เข้ามาในโปรแกรม
2. จัดการกับปัญหาด้านคุณภาพของข้อมูลทั้งหมด (Missing Values, Dirty Data, Inconsistent Format, Duplicates) ในทั้งสองไฟล์ และให้เหตุผลในการเลือกวิธีจัดการแต่ละปัญหา
3. หลังจากทำความสะอาดและเชื่อมข้อมูลแล้ว ให้สร้างรายงานสรุปภาพรวมของยอดขายและลูกค้า เช่น ยอดขายรวมทั้งหมด, จำนวนลูกค้าทั้งหมด, และจำนวนคำสั่งซื้อทั้งหมด
4. **ยอดขายรายเดือน:** วิเคราะห์แนวโน้มยอดขายรายเดือนในช่วง 3 เดือนที่ผ่านมา (มิถุนายน - สิงหาคม) และสร้าง **กราฟแท่ง (Bar Chart)** เพื่อแสดงผล
5. **แนวโน้มยอดขายรายวัน:** วิเคราะห์แนวโน้มยอดขายเฉลี่ยรายวัน และสร้าง **กราฟเส้น (Line Chart)** เพื่อแสดงการเปลี่ยนแปลงของยอดขาย
6. **สัดส่วนยอดขาย:** วิเคราะห์สัดส่วนยอดขายของแต่ละประเภทสินค้าในแต่ละเดือน และสร้าง **กราฟวงกลม (Pie Chart)** หรือ **กราฟแท่งแบบ Stacked Bar Chart** เพื่อแสดงผล
7. **พฤติกรรมการซื้อตามระดับสมาชิก:** วิเคราะห์พฤติกรรมการซื้อของลูกค้าแต่ละกลุ่มตาม membership_level (Bronze, Silver, Gold) โดยเปรียบเทียบยอดใช้จ่ายเฉลี่ยต่อคำสั่งซื้อ และสร้าง **กราฟแท่งแบบ Grouped Bar Chart** เพื่อแสดงผล
8. **การเปรียบเทียบเมือง:** ทดสอบสมมติฐานทางสถิติเพื่อหาคำตอบว่ายอดขายเฉลี่ยของลูกค้าในเมือง Bangkok แตกต่างจากเมือง Chiang Mai อย่างมีนัยสำคัญหรือไม่? (ให้ละเลยผลการตรวจสอบ Assumption)
9. **การกระจายตัวของราคา:** สร้าง **กราฟ Histogram** เพื่อแสดงการกระจายตัวของ total_price และระบุค่า Outliers ที่พบ
10. **ยอดขายตามเมืองและประเภทสินค้า:** ใช้ **Heatmap** เพื่อแสดงความสัมพันธ์ระหว่าง product_category กับ city โดยใช้ยอดขายเป็นตัวแปรหลัก
11. **ความสัมพันธ์ของข้อมูล:** สร้าง **กราฟ Scatter Plot** เพื่อแสดงความสัมพันธ์ระหว่าง total_price กับ quantity และคำนวณค่า Correlation เพื่อยืนยันความสัมพันธ์
12. **อายุการเป็นลูกค้า:** สร้างคอลัมน์ใหม่ที่ชื่อ customer_lifetime เพื่อคำนวณว่าลูกค้าแต่ละคนเป็นสมาชิกมาแล้วกี่วัน (นับจาก registration_date ถึงวันสุดท้ายที่มีการซื้อ)
13. **พฤติกรรมลูกค้าใหม่ vs. เก่า:** วิเคราะห์และสร้าง **กราฟเส้น** เพื่อเปรียบเทียบยอดขายเฉลี่ยต่อวันของลูกค้าที่สมัครสมาชิกใหม่ในแต่ละเดือนกับลูกค้าเก่า
14. **ช่วงเวลาการซื้อซ้ำ:** วิเคราะห์หาช่วงเวลาที่ลูกค้ามีแนวโน้มจะซื้อสินค้าซ้ำมากที่สุด (เช่น ซื้อซ้ำภายในกี่วัน)
15. จากการวิเคราะห์ทั้งหมดที่ทำมา น้องจะให้ข้อเสนอแนะ 3 ข้อแก่ผู้บริหารเพื่อเพิ่มยอดขายและรักษาลูกค้าได้อย่างไร? (ให้ระบุข้อมูลและกราฟที่ใช้สนับสนุนข้อเสนอแนะแต่ละข้อ)

ชุดคำถามที่ 2: การสร้าง Machine Learning Model

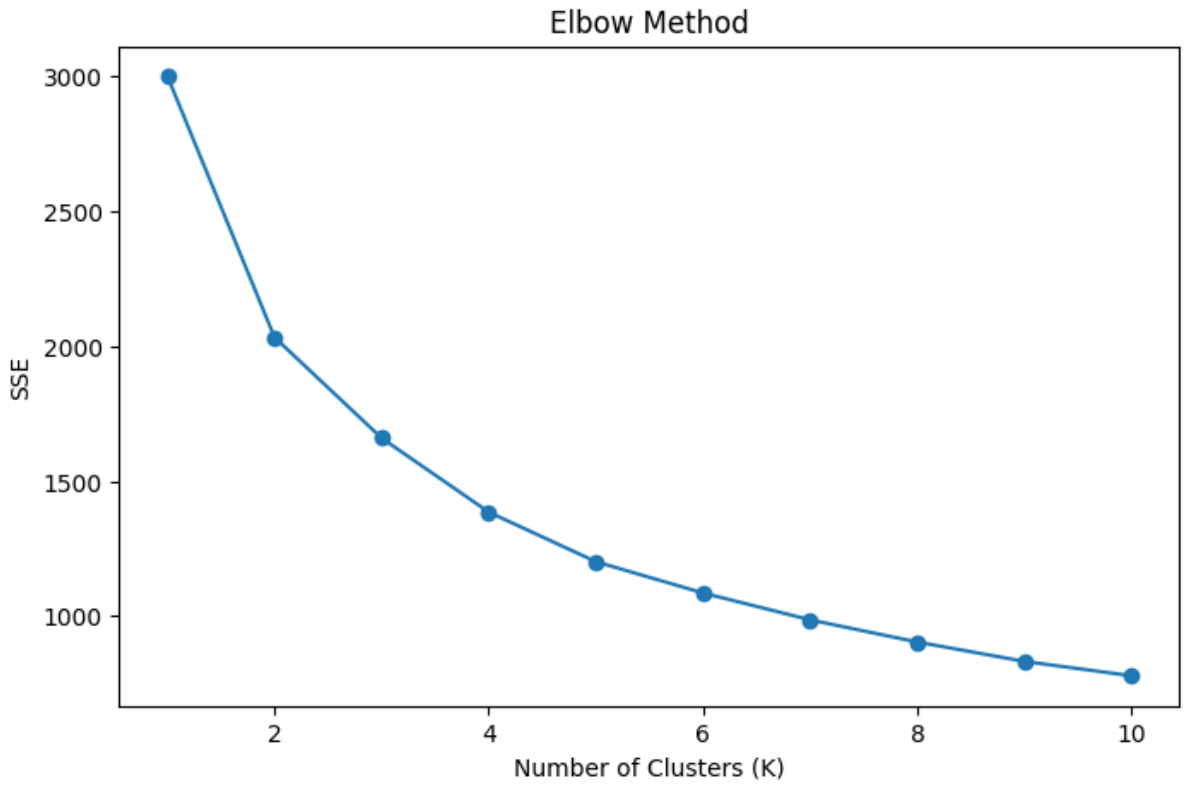
1. **การเตรียมข้อมูลสำหรับโมเดล Logistic Regression:**
    - โหลดไฟล์ ml_ready_features.csv เข้าสู่โปรแกรม Python
    - **โจทย์:** ให้สร้างคอลัมน์ใหม่ชื่อ **is_high_spender** โดยมีค่าเป็น 1 ถ้าลูกค้ามี total_spend_last_3_months มากกว่าค่าเฉลี่ยของยอดใช้จ่ายทั้งหมด และเป็น 0 ถ้าต่ำกว่า
    - เลือก Features สำหรับโมเดล (total_spend_last_3_months, purchase_count_last_3_months, avg_spend_per_purchase, customer_lifetime, membership_level)
    - จัดการกับข้อมูลประเภท Categorical ด้วย One-Hot Encoding และทำการแบ่งชุดข้อมูลเป็น Training Set และ Testing Set
2. **การสร้างและ Visualization โมเดล Logistic Regression:**
    - สร้างโมเดล **Logistic Regression** โดยใช้ scikit-learn เพื่อทำนายโอกาสที่ลูกค้าจะเป็นกลุ่มที่ใช้จ่ายสูง (High-Spender)
    - **Visualization:** เลือก Features ที่เหมาะสมมา 1-2 ตัว เช่น total_spend_last_3_months และสร้าง **กราฟ Scatter Plot** ที่แสดงความสัมพันธ์ พร้อมลากเส้น **Sigmoid Function (S-Curve)** และ **Decision Boundary** ลงบนกราฟ เพื่ออธิบายหลักการทำงานของโมเดล
3. **การประเมินผลและการตีความ Logistic Regression:**
    - ประเมินประสิทธิภาพของโมเดลด้วย Confusion Matrix, Accuracy Score, Precision, Recall, และ F1-Score
    - อธิบายว่าค่าใดสำคัญที่สุดสำหรับโจทย์นี้ และอธิบายว่าทำไม
4. **การสร้างและ Visualization โมเดล K-Means Clustering:**
    - ใช้คุณสมบัติ (total_spend_last_3_months, purchase_count_last_3_months, avg_spend_per_purchase, customer_lifetime) เพื่อจัดกลุ่มลูกค้า
    - ทำการปรับขนาดข้อมูล (Standardization หรือ Normalization) และอธิบายเหตุผลที่ต้องทำ
    - ใช้เทคนิค **Elbow Method** เพื่อหาจำนวน Cluster (K) ที่เหมาะสมที่สุด และสร้าง **กราฟ Elbow Plot** เพื่อสนับสนุนการตัดสินใจ
    - สร้างโมเดล **K-Means Clustering** และนำผลลัพธ์มาสร้าง **กราฟ Scatter Plot 2 มิติ** ที่แสดง Cluster ที่แตกต่างกัน

5. **การสรุปและข้อเสนอแนะทางธุรกิจ:**
    - เมื่อได้ Cluster ที่เหมาะสมแล้ว ให้ทำการวิเคราะห์ลักษณะของลูกค้าในแต่ละ Cluster เพื่อตั้งชื่อที่เหมาะสม (เช่น "นักช็อปประหยัด", "นักช็อปขาประจำ")
    - จากผลลัพธ์ของทั้งสองโมเดลที่ทำมา คุณจะให้ข้อเสนอแนะทางธุรกิจที่เฉพาะเจาะจงสำหรับทีมการตลาดได้อย่างไรบ้าง?
6. **การสรุปและข้อเสนอแนะทางธุรกิจ:**
    - เมื่อได้ Cluster ที่เหมาะสมแล้ว ให้ทำการวิเคราะห์ลักษณะของลูกค้าในแต่ละ Cluster (โดยเชื่อมข้อมูลกลับไปยัง membership_level และ city)
    - ตั้งชื่อให้กับแต่ละ Cluster ตามลักษณะที่พบ (เช่น "นักช็อปประหยัด", "นักช็อปขาประจำ", "ลูกค้ากระเป๋าหนัก")
    - จากผลลัพธ์ของทั้งสองโมเดล น้องจะให้ข้อเสนอแนะทางธุรกิจที่เฉพาะเจาะจงสำหรับทีมการตลาดได้อย่างไรบ้าง? เช่น "เราควรทำแคมเปญลดราคาสำหรับสินค้า Electronics ให้กับลูกค้ากลุ่ม [ชื่อ Cluster] เพราะโมเดลทำนายว่าพวกเขามีโอกาสสูงที่จะซื้อ"

หมายเหตุ



รูปที่ 1 การสร้าง Sigmoid Curve สำหรับ Logistics Regression



รูปที่ 2 ผลลัพธ์จากการทำ Elbow Method Output