สรุปทฤษฎี/เทคนิคสำคัญที่ได้เรียนมา

1. **ทำไมต้อง XAI**

   AI ทำนายเก่งแต่ไม่บอกเหตุผล = คนไม่ค่อยเชื่อและแก้ปัญหายาก โดยเฉพาะงานเสี่ยงสูง (แพทย์ การเงิน) จึงต้องมี “คำอธิบาย” ให้เห็นว่าอะไรผลักให้ได้คำตอบนั้น

2. **Interpretability vs Explainability**

   Interpretability = โมเดลอ่านโครงสร้างแล้วเข้าใจได้เลย (เช่น linear / tree)

   Explainability = โมเดล Blackbox แต่ใช้เครื่องมือช่วยอธิบายเหตุผลทีหลัง

   สรุป: Interpretability “อ่านป้ายก็รู้เรื่อง” แต่ Explainability “ต้องมีล่ามแปล”

3. **Global vs Local**

   Global = ภาพรวมทั้งโมเดล (ฟีเจอร์ไหนสำคัญโดยรวม)

   Local = รายเคส (ทำไมเคสนี้ถึงได้คำตอบแบบนี้)

4. **โมเดลตีความได้เอง vs อธิบายทีหลัง**

   ตีความได้เอง: Linear/Logistic, Decision Tree, กฎ if–then ⟶ อธิบายตรง ๆ

   อธิบายทีหลัง: Random Forest/Boosting/Neural Net ⟶ ใช้ LIME/SHAP ช่วยแปล

5. **PDP (Partial Dependence)**

   ขยับค่าฟีเจอร์ตัวเดียว แล้วดูผลทำนาย “เฉลี่ยของทั้งชุด” ⟶ เห็นความสัมพันธ์คร่าว ๆ

   ตัวอย่าง: ยิ่งอุณหภูมิสูง ยอดเช่าจักรยานอาจเพิ่มเป็นเส้นโค้งขึ้น

6. **ICE (Individual Conditional Expectation)**

   เหมือน PDP แต่แยกเป็น “เส้นต่อคน/ต่อเคส” ⟶ เห็นว่าแต่ละเคสอาจตอบสนองต่างกัน

   ตัวอย่าง: คนที่แพ้ร้อนอาจเช่าน้อยลง แม้อุณหภูมิสูง (เส้นของเคสนี้จะไม่เหมือนค่าเฉลี่ย)

7. **LIME**

   ทำข้อมูล “คล้าย ๆ เคสเป้าหมาย” หลายแบบ แล้วให้โมเดลตอบ จากนั้นใช้โมเดลง่าย ๆ ใกล้จุดนั้น อธิบายว่า ฟีเจอร์ไหนดันผลขึ้น/ลง

8. **SHAP**

   แบ่ง “เครดิต” ความสำคัญของฟีเจอร์ต่อผลทำนายรายเคส ด้วยหลัก Shapley ทำให้อ่านได้ทั้งภาพรวมและรายเคส

   สรุป: “แบ่งแต้มให้ฟีเจอร์อย่างเป็นธรรมว่าใครช่วยให้คะแนนสุดท้ายเป็นเท่าไหร่”

9. **Grad-CAM**

   ทำฮีตแมปว่าบริเวณไหนในรูปมีผลต่อคำตอบมาก ⟶ เช็กได้ว่าโมเดลมอง “ตัววัตถุ” ไม่ใช่ฉากหลัง

   ตัวอย่างเช่น: ทำนาย “แมว” แล้วแผนที่ความร้อนขึ้นที่หน้าหูตาแมว ไม่ใช่มุมพื้น

10. **Counterfactual Explanation**

"ถ้าเปลี่ยนเล็กน้อย ผลจะเปลี่ยนไหม"

ตัวอย่างเช่น: "ถ้าเงินเดือนเพิ่มอีก 3,000 บาท จะผ่านอนุมัติสินเชื่อไหม" บอกแนวทางปรับจริงได้

11. Logistic Regression

เป็นโมเดลแบบ "รวมคะแนนแล้วแปลงเป็นโอกาส" แต่ละปัจจัยมีคะแนนบวกหรือลบ ถ้าปัจจัยไหนช่วยเพิ่มโอกาส คะแนนนั้นเป็นบวก ถ้าลดโอกาสเป็นลบ รวมกันแล้วได้ตัวเลข 0–1 บอกว่า "น่าจะใช่แค่ไหน"

12. **Decision Tree**

เป็นโมเดลที่ตอบคำถามใช่/ไม่ใช่ทีละข้อ เช่น "อายุเกิน 50 ไหม?" ถ้าใช่ไปทางซ้าย ถ้าไม่ใช่ไปทางขวา เดินไปเรื่อย ๆ จนได้คำตอบสุดท้าย

13. **EBM (Explainable Boosting Machine)**

เป็นโมเดลแบบรวม "เส้นโค้งผลกระทบ" ต่อฟีเจอร์เป็นผลรวมเพื่อทำนาย; เห็นชัดว่าค่าฟีเจอร์เพิ่ม/ลดส่งผลอย่างไร ให้ความแม่นยำสูงพร้อมความอธิบายง่าย

14. Forest-Guided Clustering

เป็นการเอาโครงสร้างของ Random Forest มาช่วยจับกลุ่มตัวอย่างที่ตัดสินใจคล้ายกันในต้นไม้หลายๆ ต้น แล้วสรุปว่าแต่ละกลุ่มเด่นเรื่องฟีเจอร์อะไร ผลลัพธ์คือเห็นภาพรวมว่าโมเดลแบ่งข้อมูลเป็นพวก ๆ ด้วยเหตุผลแบบไหน และจุดไหนดูแปลกไปจากเพื่อนร่วมกลุ่ม

15. **Grad-CAM**

เป็น "แผนที่สี" ซ้อนบนรูปเพื่อบอกว่าโมเดลใช้ส่วนไหนของภาพในการตัดสินใจ ยิ่งสีเข้ม = มีผลมา, ยิ่งสีจาง = มีผลน้อย

ตัวอย่างเช่น ถ้าโมเดลทายว่าเป็นแมวแล้วสีร้อนไปกองที่หน้า แปลว่าโมเดลมองถูกจุด แต่ถ้าสีไปรวมที่ฉากหลังแปลว่าโมเดลอาจมองผิดจุด

16. ทำไมต้อง XAI ใน NLP

โมเดลภาษาจะอ่านคำหลายๆคำต่อกัน เราอยากรู้ว่า "คำไหนหรือช่วงไหน" ทำให้ได้คำตอบ เพื่อเช็กว่าคิดถูกจุดและอธิบายให้ผู้ใช้เข้าใจ

- **LIME สำหรับข้อความ**

  แนวคิดคือ ลองปิดคำทีละคำแล้วดูว่าใจความ/คำตอบเปลี่ยนไหม ถ้าเปลี่ยน แปลว่าคำนั้นมีอิทธิพลมาก เหมือนลองเอาชิ้นส่วนหนึ่งออกจากประโยคแล้วดูว่ายังเหมือนเดิมไหม

- **SHAP สำหรับข้อความ**

  คือวิธีอธิบายผลทำนายที่แบ่งแต้มให้แต่ละฟีเจอร์ ค่าเป็นบวกแปลว่าช่วยดันผลไปทางคำตอบนั้น ค่าเป็นลบแปลว่าฉุดออกห่าง ค่ายิ่งใหญ่ยิ่งมีอิทธิพล และเมื่อ รวมแต้มทุกฟีเจอร์จะเท่ากับผลทำนาย

17. Gradient-based Explanations

การสร้าง Gradient-based จะมองไปที่ map ของการเชื่อมโยง เช่นการมองภาพของคำที่เรียงต่อกัน ถ้าเรียง X-X-X-X แปลว่า XXXX การที่เราเอาลำดับมาใช้จะเรียกว่า Gradient-based
ตัวอย่างเช่น จากการที่ก่อนหน้านี้เรามองว่า ฟีเจอร์นี้มีผลต่อโมเดล แต่ตอนนี้เราจะมองภาพเพิ่มเป็น Gradient-based คือ ฟีเจอร์นี้และฟีเจอร์นี้ เรียงต่อกันมีผลต่อโมเดล โดยกระบวนการนี้มีเครื่องมือที่ชื่อว่า saliency map

18. Attention

Attention บอกว่าโมเดลโฟกัสตรงไหน
ตัวอย่างเช่น หากเราต้องการที่จะเข้าใจประโยคนี้ของข้อความ เราหาเฉพาะ Attention จุดเด่น ๆ ในข้อความออกมา แล้วบอกว่านี่คือตัวแทน แล้วเราบอกว่านี่คือความหมาย แต่ Attention ก็ไม่ได้เป็นความหมายของข้อความเสมอไปเพราะบาง Attention ไม่ได้แสดงถึงความหมายของข้อความ แค่ Attention ตัวนั้นีความถี่เยอะ

19. บทเรียนด้านจริยธรรมและ fairness

คิดเรื่องนี้ตั้งแต่ต้นจนจบงานเสมอ เก็บข้อมูลให้ครบทุกกลุ่ม ไม่เอนเอียง, เลือกตัวแปรที่ไม่พาไปแตะประเด็นอ่อนไหวทางอ้อม เช่น เลี่ยงใช้รหัสไปรษณีย์แทนเชื้อชาติ, ตอนวัดผลอย่าดูแค่ความแม่น ให้ดูด้วยว่าผลลัพธ์ยุติธรรมพอ ๆ กันในแต่ละกลุ่มหรือไม่, และอธิบายผลให้เข้าใจง่าย พร้อมทั้งทำให้แก้ไขได้