PROJECT PROPOSAL

รายละเอียดโครงการฝึกงาน / สหกิจศึกษา

Return & Refund Risk Scoring System

ระบบทำนายความเสี่ยงการคืนสินค้าสำหรับ GMM O Shopping

บริษัท / สถานที่ฝึกงาน	GMM O Shopping (gmmoshopping.com)
ประเภทโปรเจ็กต์	Data Science / Machine Learning
ระยะเวลาดำเนินงาน	2 เดือน (8 สัปดาห์)
เครื่องมือหลัก	Python, SQL, XGBoost, SHAP, Streamlit
เป้าหมายธุรกิจ	ลดต้นทุน Logistics จากการคืนสินค้า

1. แนวคิดของโปรเจ็กต์ (Concept)

1.1 ที่มาและปัญหา

GMM O Shopping เป็นธุรกิจ Home Shopping E-Commerce ที่มีช่องทางขายสินค้าหลายรูปแบบ ทั้งโทรทัศน์ตลอด 24 ชั่วโมง, เว็บไซต์, แอปพลิเคชัน และ Social Commerce ด้วยลักษณะของธุรกิจที่ลูกค้าตัดสินใจซื้อสินค้าจากการนำเสนอผ่านหน้าจอโดยไม่ได้จับหรือทดลองสินค้าจริงก่อน ส่งผลให้อัตราการคืนสินค้า (Return Rate) ของ Home Shopping สูงกว่า E-Commerce ทั่วไปอย่างมีนัยสำคัญ

การคืนสินค้าแต่ละครั้งก่อให้เกิดต้นทุนที่ซ่อนอยู่หลายด้าน ได้แก่ ค่าขนส่งสินค้ากลับ ค่าตรวจสอบและบรรจุสินค้าใหม่ ค่าจัดเก็บสินค้าที่รอตรวจสอบ และมูลค่าสินค้าที่เสื่อมลงระหว่างกระบวนการคืน รวมถึงภาระงานของทีม Customer Service ที่ต้องดูแลกระบวนการคืนสินค้าในแต่ละเคส

1.2 แนวคิดหลัก

โปรเจ็กต์นี้ตั้งอยู่บนแนวคิดที่ว่า "การป้องกันการคืนสินค้าที่ดีที่สุดคือการรู้ล่วงหน้าว่าออเดอร์ไหนจะถูกคืน" แทนที่จะรอให้ลูกค้าแจ้งคืนสินค้าแล้วค่อยดำเนินการ ระบบนี้จะทำการประเมินความเสี่ยงของออเดอร์ตั้งแต่วินาทีที่ลูกค้ากดสั่งซื้อ

แนวคิดนี้อาศัยหลักการของ Predictive Analytics คือการใช้ข้อมูลในอดีตของบริษัท เช่น ประวัติออเดอร์ที่เคยถูกคืน ลักษณะของลูกค้า ประเภทสินค้า ช่องทางการซื้อ และปัจจัยอื่นๆ มาฝึก Machine Learning Model ให้เรียนรู้รูปแบบ (Pattern) ของออเดอร์ที่มีแนวโน้มจะถูกคืน จากนั้น Model จะทำการประเมิน Risk Score ให้กับออเดอร์ใหม่ทุกรายการแบบอัตโนมัติ

1.3 สิ่งที่ทำให้โปรเจ็กต์นี้แตกต่าง

- ใช้ข้อมูลจริงของ GMM O Shopping ซึ่งมีลักษณะเฉพาะของ Home Shopping ที่ไม่มีใน Dataset สาธารณะ
- รวม Feature จากหลายมิติ ทั้งข้อมูลลูกค้า สินค้า ช่องทางการซื้อ และพฤติกรรมในอดีต
- ใช้ Cost-sensitive Learning คือไม่ได้แค่ maximize accuracy แต่คำนึงถึงต้นทุนจริงของแต่ละประเภทของ Error
- มี SHAP Explainability ทำให้ทีม Business เข้าใจได้ว่าทำไมออเดอร์นั้นถึงได้ Risk Score สูง

2. หลักการทำงาน (How It Works)

2.1 Machine Learning Classification

ระบบนี้ใช้หลักการของ Binary Classification กล่าวคือ Model จะทำการเรียนรู้จากข้อมูลออเดอร์ในอดีต โดยแต่ละออเดอร์จะมี Label ว่า "ถูกคืน" หรือ "ไม่ถูกคืน" จากนั้นเมื่อมีออเดอร์ใหม่เข้ามา Model จะประเมินว่าออเดอร์นั้นมีความน่าจะเป็น (Probability) ที่จะถูกคืนสูงแค่ไหน โดยแสดงผลเป็น Risk Score ในระดับ 0.0 ถึง 1.0

2.2 Feature ที่ Model ใช้ในการประเมิน

Model จะประเมินจาก 4 กลุ่ม Feature หลัก ดังนี้

กลุ่มที่ 1: ข้อมูลลูกค้า (Customer Features)

- ประวัติการคืนสินค้าในอดีต — ลูกค้าที่เคยคืนสินค้าบ่อยมีแนวโน้มคืนอีก
- จำนวนครั้งที่ซื้อทั้งหมด (Purchase Frequency)
- อายุบัญชีของลูกค้า (Days Since Registration)
- ระดับสมาชิก (Membership Tier)
- ช่องทางที่ลูกค้านิยมซื้อ (Preferred Channel)

กลุ่มที่ 2: ข้อมูลสินค้า (Product Features)

- หมวดหมู่สินค้า — บางหมวดมี Return Rate สูงกว่าหมวดอื่น
- ราคาสินค้า — สินค้าราคาสูงมักถูกพิจารณาคืนมากกว่า
- Return Rate เฉลี่ยของสินค้า SKU นั้นในอดีต
- Brand และ Supplier ของสินค้า

กลุ่มที่ 3: ข้อมูลออเดอร์ (Order Features)

- ช่องทางที่ซื้อ เช่น TV / เว็บไซต์ / แอป / โทรศัพท์
- โปรโมชันหรือส่วนลดที่ใช้ (Discount Percentage)
- วิธีชำระเงิน เช่น บัตรเครดิต / โอนเงิน / เก็บเงินปลายทาง
- วันและเวลาที่สั่งซื้อ
- จำนวนสินค้าในออเดอร์

กลุ่มที่ 4: Feature ที่สร้างขึ้นใหม่ (Engineered Features)

- return_rate_by_category — อัตราการคืนเฉลี่ยของสินค้าประเภทเดียวกัน
- customer_return_ratio — สัดส่วนออเดอร์ที่ลูกค้าเคยคืนต่อออเดอร์ทั้งหมด
- days_since_last_return — ระยะเวลาตั้งแต่คืนสินค้าครั้งล่าสุด
- high_risk_hour — ชั่วโมงที่ผิดปกติซึ่งสัมพันธ์กับการคืน

2.3 Algorithm ที่ใช้

โปรเจ็กต์จะเปรียบเทียบ Model หลายตัว โดย Model หลักที่คาดว่าจะให้ผลดีที่สุดคือ XGBoost (Extreme Gradient Boosting) ซึ่งเป็น Algorithm ที่ได้รับความนิยมสูงในงาน Tabular Data ประกอบด้วย

- Logistic Regression — ใช้เป็น Baseline เพื่อเปรียบเทียบ
- Random Forest — ทนทานต่อ Noise และ Overfitting
- XGBoost / LightGBM — คาดว่าให้ผลดีที่สุด เร็วและแม่นยำ

เนื่องจากจำนวนออเดอร์ที่ถูกคืนมักน้อยกว่าออเดอร์ปกติมาก (Imbalanced Data) จะใช้เทคนิค SMOTE และ Class Weight Adjustment เพื่อแก้ปัญหานี้ และใช้ Cost Matrix ในการ Tune Model ให้ระวังการพลาด High-risk Order มากกว่าการ False Alarm

3. Workflow การดำเนินงาน

กระบวนการทำงานของโปรเจ็กต์แบ่งออกเป็น 5 ขั้นตอนหลัก ตามกรอบ CRISP-DM (Cross-Industry Standard Process for Data Mining)

#	ขั้นตอน	รายละเอียด	Output
1	**Data Collection & Understanding**	ดึงข้อมูล Order, Return, Customer จาก Database ด้วย SQL ศึกษา Schema และตรวจสอบ Data Quality เบื้องต้น	Data Dictionary, Clean Dataset
2	**Exploratory Data Analysis (EDA)**	วิเคราะห์ Pattern ของการคืนสินค้า หา Correlation ระหว่าง Feature และ Return Label สร้าง Visualization แสดง Business Insight	EDA Report, Visualization
3	**Feature Engineering & Preprocessing**	สร้าง Feature ใหม่จากข้อมูลดิบ แปลง Categorical Variables จัดการ Missing Values แบ่ง Train/Test Set จัดการ Imbalanced Data	Feature Set พร้อม Train
4	**Model Training & Evaluation**	ฝึก Model หลายตัว Tune Hyperparameter ด้วย Optuna ประเมินผลด้วย Cost Matrix และ AUC-ROC วิเคราะห์ด้วย SHAP	Best Model (.pkl), Metrics Report
5	**Dashboard & Reporting**	สร้าง Interactive Dashboard แสดง Risk Score รายออเดอร์ แบ่ง Tier Low/Medium/High เขียน Final Report ประมาณการ Cost Saving	Dashboard, Final Report

3.1 รายละเอียด Workflow ขั้นตอนที่ 1 — Data Collection

เริ่มต้นด้วยการเขียน SQL Query เพื่อดึงข้อมูลจาก Database ของบริษัท โดยจะ JOIN ตาราง Orders, Returns, Customers และ Products เข้าด้วยกัน จากนั้นนำข้อมูลมาตรวจสอบด้วย Python เพื่อหา Missing Values, Duplicate Records และ Outliers พร้อมทำ Data Dictionary บันทึกความหมายของแต่ละ Column

3.2 รายละเอียด Workflow ขั้นตอนที่ 2 — EDA

ใช้ matplotlib และ seaborn สร้าง Visualization เพื่อตอบคำถามสำคัญ เช่น หมวดสินค้าไหนมี Return Rate สูงสุด?, ช่องทางที่ซื้อมีผลต่อการคืนสินค้าอย่างไร?, ลูกค้ากลุ่มไหนที่มีแนวโน้มคืนบ่อยที่สุด? และ ช่วงเวลาหรือโปรโมชันไหนที่สัมพันธ์กับการคืน?

3.3 รายละเอียด Workflow ขั้นตอนที่ 3 — Feature Engineering

ขั้นตอนนี้คือการ "สร้างข้อมูลใหม่จากข้อมูลเดิม" เช่น การคำนวณ Return Rate เฉลี่ยของสินค้าแต่ละ SKU ในช่วง 3 เดือนที่ผ่านมา หรือการสร้าง Feature ว่าลูกค้าคนนี้เคยคืนสินค้ากี่ครั้งใน 6 เดือนล่าสุด ซึ่ง Feature เหล่านี้มักมีพลังในการ Predict สูงกว่า Feature ดิบ

3.4 รายละเอียด Workflow ขั้นตอนที่ 4 — Model Training

ใช้ scikit-learn pipeline ฝึก Model หลายตัว จากนั้นใช้ Optuna หา Hyperparameter ที่ดีที่สุดอัตโนมัติ เมื่อได้ Best Model แล้ว จะใช้ SHAP (SHapley Additive exPlanations) วิเคราะห์ว่า Feature ไหนมีผลต่อการตัดสินใจของ Model มากที่สุด และอธิบายได้ในระดับรายออเดอร์ว่าทำไมออเดอร์นี้ถึงได้ Risk Score สูง

3.5 รายละเอียด Workflow ขั้นตอนที่ 5 — Dashboard

สร้าง Dashboard ด้วย Streamlit ที่แสดงรายการออเดอร์พร้อม Risk Score และ Risk Tier (Low / Medium / High) ทีม Operations สามารถ Filter ออเดอร์ High-risk ออกมาและดำเนินการเชิงรุกได้ทันที Dashboard จะแสดง SHAP Explanation ของแต่ละออเดอร์ด้วย เพื่อให้ทีมเข้าใจว่าเหตุใดออเดอร์นั้นถึงมีความเสี่ยงสูง

4. การวัดผลและ Output ของระบบ

4.1 Metrics ที่ใช้วัดผล Model

โปรเจ็กต์นี้ไม่ได้ใช้แค่ Accuracy เพียงอย่างเดียว เนื่องจากข้อมูลมีความไม่สมดุล (Imbalanced) จึงใช้ Metrics หลายตัวประกอบกัน โดยเฉพาะอย่างยิ่ง Recall ซึ่งวัดว่าระบบสามารถจับ High-risk Order ได้มากแค่ไหน

Metric	ความหมาย	เป้าหมาย
Precision	ในบรรดาที่ระบบบอกว่าเสี่ยง มีที่เสี่ยงจริงกี่ %	> 70%
Recall (สำคัญที่สุด)	ในบรรดาที่เสี่ยงจริง ระบบจับได้กี่ %	> 75%
F1-Score	ค่าเฉลี่ยของ Precision และ Recall	> 72%
AUC-ROC	ความสามารถแยกแยะ High/Low risk	> 0.80
Cost Matrix Score	ต้นทุนรวมที่ลดได้ถ้าใช้ระบบจริง	ลดต้นทุน > 20%

4.2 Output ที่ได้จากระบบ

Output 1: Risk Score รายออเดอร์

ออเดอร์ทุกรายการจะได้รับ Risk Score เป็นตัวเลขระหว่าง 0.0 ถึง 1.0 โดยแบ่งเป็น 3 ระดับ ดังนี้

Risk Tier	Risk Score	Action ที่ควรดำเนินการ
Low Risk	0.0 – 0.3	ดำเนินการปกติ ส่งสินค้าตามขั้นตอนมาตรฐาน
Medium Risk	0.3 – 0.6	ส่ง Follow-up Email / SMS หลังรับสินค้า ตรวจสอบความพึงพอใจ
High Risk	0.6 – 1.0	โทรยืนยันก่อนส่ง / บรรจุภัณฑ์พิเศษ / แนบคู่มือการใช้งานเพิ่ม

Output 2: SHAP Explanation รายออเดอร์

สำหรับออเดอร์ High-risk ทุกรายการ ระบบจะแสดงคำอธิบายว่า Feature ใดที่ส่งผลให้ Risk Score สูง เช่น "ลูกค้ารายนี้เคยคืนสินค้า 3 ครั้งใน 6 เดือนที่ผ่านมา" หรือ "สินค้าหมวดนี้มี Return Rate ถึง 25%" ซึ่งช่วยให้ทีม Operations ตัดสินใจได้ว่าควร intervene อย่างไร

Output 3: Dashboard สรุปภาพรวม

นอกจาก Risk Score รายออเดอร์แล้ว Dashboard ยังแสดงภาพรวมของระบบ

เช่น จำนวนออเดอร์ High-risk ในวันนี้, หมวดสินค้าที่มี Risk สูงที่สุด, แนวโน้ม Return Rate รายสัปดาห์ และประมาณการต้นทุน Logistics ที่ประหยัดได้จากการใช้ระบบ

Output 4: Final Report พร้อมประมาณการ ROI

เมื่อสิ้นสุดโปรเจ็กต์จะจัดทำรายงานสรุปที่ประเมิน Return on Investment (ROI) ของระบบ โดยคำนวณจากจำนวน High-risk Order ที่ระบบตรวจพบได้คูณกับต้นทุนการคืนสินค้าเฉลี่ยต่อออเดอร์ เพื่อแสดงให้เห็นว่าถ้านำระบบไปใช้จริงจะประหยัดต้นทุนได้เท่าไร

5. ประโยชน์ต่อธุรกิจ GMM O Shopping

5.1 ประโยชน์โดยตรงต่อแต่ละทีม

	ทีม / ผู้ที่ได้รับประโยชน์	ประโยชน์หลัก	รายละเอียด
OPS	**ทีม Operations**	**ลดภาระงาน Return**	รู้ล่วงหน้าว่าออเดอร์ไหนเสี่ยง ไม่ต้องรอลูกค้าแจ้งคืน สามารถ intervene ได้ก่อน
CRM	**ทีม CRM / Care**	**Proactive Service**	โทรหาลูกค้า High-risk ก่อนส่งสินค้า เพิ่มความพึงพอใจและลด Churn Rate
FIN	**ทีม Finance**	**ลดต้นทุน Logistics**	ค่าขนส่งคืน + ค่าตรวจสอบสินค้าลดลง โดยเฉพาะ SKU ที่มี Return Rate สูงเป็นประจำ
MKT	**ทีม Marketing**	**ปรับ Campaign**	รู้ว่าโปรโมชันไหนทำให้ Return Rate สูง สามารถปรับ Mechanic โปรโมชันได้
BUY	**ทีม Buyer**	**ปรับ Sourcing**	รู้ว่าสินค้า Supplier ไหนมี Return สูง สามารถเจรจาเงื่อนไขหรือเปลี่ยน Supplier ได้

5.2 ผลกระทบเชิงตัวเลขที่คาดหวัง

สมมติว่า GMM O Shopping มีออเดอร์เฉลี่ย 10,000 รายการต่อเดือน และมี Return Rate 8% (800 ออเดอร์) โดยต้นทุนการคืนสินค้าต่อออเดอร์เฉลี่ยอยู่ที่ 300 บาท หากระบบสามารถป้องกัน Return ได้ 30% ของ High-risk Orders จะ

ประหยัดต้นทุนได้ดังนี้

รายการ	ตัวเลขประมาณการ
Return Orders ต่อเดือน (สมมติ 8%)	**800 ออเดอร์**
ต้นทุนต่อการคืน 1 ออเดอร์	**300 บาท**
ต้นทุน Return รวมต่อเดือน	**240,000 บาท**
ถ้าระบบป้องกันได้ 30% ของ High-risk	**ลด Return ได้ ~240 ออเดอร์**
ต้นทุนที่ประหยัดได้ต่อเดือน	**72,000 บาท**
ต้นทุนที่ประหยัดได้ต่อปี (ประมาณ)	**864,000 บาท**

หมายเหตุ: ตัวเลขข้างต้นเป็นการประมาณการเพื่อแสดง Business Case เท่านั้น ตัวเลขจริงจะถูกคำนวณจากข้อมูลจริงของบริษัทในช่วงฝึกงาน

5.3 ประโยชน์ระยะยาว (ต่อยอดเป็นโปรเจ็กต์จบ)

- ต่อยอดเป็น Real-time API ที่ประเมิน Risk Score ทันทีที่ลูกค้ากดยืนยันออเดอร์ โดยไม่ต้องรัน batch process
- เพิ่ม Automated Alert System ที่แจ้งเตือนทีม Customer Care อัตโนมัติผ่าน LINE Notify หรือ Email
- วัดผลจริงในเชิง Before-After Analysis เพื่อยืนยันว่าระบบช่วยลด Return Rate ได้จริง
- เป็น Academic Contribution ในฐานะงานวิจัยแรกที่ศึกษา Return Risk Prediction สำหรับ Home Shopping ในบริบทของประเทศไทย

6. แผนการดำเนินงาน 8 สัปดาห์

สัปดาห์	หัวข้อ	รายละเอียด	เครื่องมือ
Week 1	**Business & Data Understanding**	ศึกษา DB Schema, จัดทำ Data Dictionary, ER Diagram	SQL, Draw.io
Week 2	**Data Extraction & Cleaning**	ดึงข้อมูล, จัดการ Missing/Outlier, สร้าง Clean Dataset	pandas, SQL
Week 3	**Exploratory Data Analysis**	วิเคราะห์ Pattern, สร้าง Visualization, สรุป Insight	matplotlib, seaborn
Week 4	**Feature Engineering**	สร้าง Feature ใหม่, Encode Categorical, แบ่ง Train/Test	pandas, scikit-learn

Week 5	**Model Training & Tuning**	Train 3 Models, จัดการ Imbalanced Data, Tune HPs	XGBoost, Optuna
Week 6	**Evaluation & SHAP**	ประเมินด้วย Cost Matrix, วิเคราะห์ SHAP	SHAP, sklearn
Week 7	**Dashboard Development**	สร้าง Risk Score Dashboard, แบ่ง Risk Tier	Streamlit, Plotly
Week 8	**Final Report & Presentation**	สรุป Report, ประมาณการ Cost Saving, จัดระเบียบ GitHub	GitHub, PowerPoint