การวิเคราะห์พฤติกรรมของลูกค้าที่มีผลต่อการซื้อและการสร้างแบบจำลองการคาดการณ์

Analysis of Customer Behavior Affecting Purchase and Development of Predictive Modeling

นายวุฒิพงษ์ พุ่มประดับ 6404800023

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์

มหาวิทยาลัยสยาม

ปีการศึกษา 2567

หัวข้อปริญญานิพนธ์	การวิเคราะห์พฤติกรรมของลูกค้าที่มีผลต่อการซื้อและการสร้างแบบจำลองการคาดการณ์ Analysis of Customer Behavior Affecting Purchase and Development of Predictive Modeling
หน่วยกิตของปริญญานิพนธ์	3 หน่วยกิต
รายชื่อผู้จัดทำ	นายวุฒิพงษ์ พุ่มประดับ 6404800023
อาจารย์ที่ปรึกษา	อาจารย์จรรยา แหนมเจริญ
ระดับการศึกษา	ปริญญาตรี
ภาควิชา	วิทยาการคอมพิวเตอร์
ปีการศึกษา	2567

อนุมัติให้ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

คณะกรรมการสอบปริญญานิพนธ์

.....ประธานกรรมการ
(พล.อ.ท.ผศ.ดร. พาห์รณ สงวนโภคัย)

.....กรรมการ
(อาจารย์ธนาภรณ์ รอดชีวิต)

.....อาจารย์ที่ปรึกษา
(อาจารย์จรรยา แหยมเจริญ)

หัวข้อปริญญานิพนธ์	การวิเคราะห์พฤติกรรมของลูกค้าที่มีผลต่อการซื้อและการสร้างแบบจำลองการคาดการณ์
หน่วยกิตของปริญญานิพนธ์	3 หน่วยกิต
ผู้จัดทำ	นายวุฒิพงษ์ พุ่มประดับ 6404800023
อาจารย์ที่ปรึกษา	อาจารย์จรรยา แหยมเจริญ
ระดับการศึกษา	วิทยาศาสตรบัณฑิต
ภาควิชา	วิทยาการคอมพิวเตอร์
ปีการศึกษา	2567

บทคัดย่อ

วัตถุประสงค์ในการจัดทำปริญญานิพนธ์นี้เพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการตัดสินใจซื้อของลูกค้า (Drop Lead) โดยใช้ข้อมูลพฤติกรรมของลูกค้า ประกอบด้วย การคลิกของลูกค้าในเว็บไซต์ ในการศึกษา 1) ปัจจัยใดบ้างที่มีผลต่อการตัดสินใจซื้อของลูกค้า 2) โมเดลที่สามารถทำนายโอกาสในการตัดสินใจซื้อของลูกค้าได้อย่างแม่นยำ ขั้นตอนในการดำเนินการวิเคราะห์ข้อมูลประกอบด้วย 1) ศึกษาและทำความเข้าใจข้อมูล 2) กำหนดเป้าหมายในการวิเคราะห์ข้อมูล 3) จัดเตรียมข้อมูล 4) วิเคราะห์ข้อมูล 5) สร้างและประเมินโมเดลการทำนาย เครื่องมือในการวิเคราะห์ข้อมูล ได้แก่ โปรแกรม Jupyter Notebook เขียนชุดคำสั่งด้วยภาษาไพธอนสำหรับการเตรียมข้อมูล การวิเคราะห์พฤติกรรมของลูกค้า และการสร้างโมลเดลการทำนาย ผลลัพธ์ที่ได้จากการวิเคราะห์พฤติกรรมของลูกค้าพบว่าลูกค้าที่มีการติดต่อกับพนักงานมักจะตัดสินใจซื้อทรัพย์สิน และการแบบจำลองการพยากรณ์ Random Forest โดยที่ให้ค่า Accuracy > 99%, Recall ค่อนข้างต่ำแต่มีค่าสูงสุด และมีค่า Precision มากที่สุด บริษัทสามารถนำข้อมูลเชิงลึกที่ได้ไปวางแผนกลยุทธ์ทางการตลาดและการขายได้

คำสำคัญ : การวิเคราะห์ข้อมูล, พฤติกรรมผู้บริโภค, แบบจำลองการพยากรณ์

Project title : Analysis of Customer Behavior Affecting Purchase and Development of Predictive Modeling

Credits : 3 Units

By : Mr. Wutthipong Poompadub 6404800023

Advisor : Miss Janya Yamchareon

Program : Bachelor of Science

Major : Computer Science

Faculty : Science

Semester/Academic year : 1/2024

Abstract

The objective of this project is to analyze the factors influencing customer purchase decisions (Drop Lead) using customer behavior data, particularly customer clicks on the website. The study focuses on identifying the factors that affect customer purchase decisions and developing a model that can accurately predict the likelihood of a customer making a purchase. The data analysis process consists of the following steps: 1) studying and understanding the data; 2) defining the objectives for data analysis; 3) preparing the data; 4) analyzing the data; and 5) building and evaluating the predictive model. The tools used for data analysis includes Jupyter Notebook, and Python for data preparation, customer behavior analysis, and predictive model development. The results from the customer behavior analysis indicate that customers who interact with sales representatives are more likely to make a purchase. The predictive modeling using Random Forest achieved an accuracy of over 99%, a relatively low recall but with the highest possible value, and the highest precision. The company can utilize these insights to develop marketing and sales strategies effectively.

Keywords: data analytics, customer behaviors, predictive model

.....

(Project Advisor)

Approved by

.....

กิตติกรรมประกาศ

(Acknowledgement)

การจัดทำปริญญานิพนธ์ฉบับนี้สำเร็จได้นั้น ผู้จัดทำได้รับความกรุณาจากอาจารย์ผู้สอนทุกท่านที่ให้ข้อมูลต่างๆ ส่งผลให้ผู้จัดทำได้รับความรู้และประสบการณ์ต่างๆ ที่มีค่ามากมาย สำหรับปริญญานิพนธ์ฉบับนี้สำเร็จลงได้ด้วยดีจากความร่วมมือและสนับสนุนจากหลายฝ่าย ดังนี้

1. อาจารย์จรรยา แหยมเจริญ
2. คุณอนันต์ ตีระบูรณะพงษ์

ผู้จัดทำใคร่ขอขอบพระคุณคณะกรรมการสอบปริญญานิพนธ์ ที่ได้ให้คำแนะนำสำคัญในการสอบปริญญานิพนธ์ฉบับนี้ และผู้มีส่วนร่วมทุกท่าน รวมทั้งผู้ที่ไม่ได้กล่าวนามที่มีส่วนร่วมในการให้ข้อมูลให้ความช่วยเหลือ และเป็นที่ปรึกษาให้คำแนะนำต่างๆ จนทำให้งานทุกอย่างประสบความสำเร็จไปด้วยดี และจัดทำรายงานฉบับนี้จนเสร็จสมบูรณ์ ซึ่งผู้จัดทำขอขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้ด้วย

ผู้จัดทำ

นาย วุฒิพงษ์ พุ่มประดับ

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบัน การทำความเข้าใจพฤติกรรมของลูกค้าและการคาดการณ์พฤติกรรมในอนาคตได้กลายเป็นส่วนสำคัญในการเพิ่มศักยภาพทางธุรกิจ โดยเฉพาะในอุตสาหกรรมอสังหาริมทรัพย์ที่ต้องเผชิญกับการแข่งขันอย่างเข้มข้น การวิเคราะห์ข้อมูลพฤติกรรมของผู้ใช้และลูกค้า เช่น การเข้าชมเว็บไซต์ การคลิกสอบถามข้อมูล การเปรียบเทียบสินค้า และการตัดสินใจซื้อ เป็นเครื่องมือที่ช่วยให้บริษัทสามารถเข้าใจแนวโน้มและความต้องการของลูกค้าในเชิงลึก

นอกจากนี้ การสร้างแบบจำลองการพยากรณ์ (Predictive Model) ยังเป็นอีกหนึ่งวิธีสำคัญที่ช่วยให้บริษัทสามารถคาดการณ์ได้ว่าลูกค้าที่มีพฤติกรรมลักษณะใดจึงจะมีแนวโน้มที่จะตัดสินใจซื้อสินค้า โมเดลนี้ไม่เพียงแต่ช่วยลดการสูญเสียโอกาสทางการขาย (Drop lead) แต่ยังช่วยเพิ่มโอกาสในการตอบสนองต่อความต้องการของลูกค้าอย่างตรงจุด

จากที่กล่าวมาข้างต้นทางผู้จัดทำ จึงได้เล็งเห็นถึงปัญหาถึงพฤติกรรมใดของลูกค้าที่จะมีผลต่อการตัดสินใจซื้อสินค้า (Drop lead) จากนั้นจึงทำการวิเคราะห์พฤติกรรมของลูกค้าโดยใช้ข้อมูลพฤติกรรมการซื้อของลูกค้าจากบริษัทสินทรัพย์แห่งหนึ่ง เพื่อใช้ในการหาพฤติกรรมที่จะนำไปสู่การตัดสินใจซื้อทรัพย์สิน ขั้นตอนในการดำเนินการวิเคราะห์ประกอบด้วย 1) ศึกษาและทำความเข้าใจข้อมูล 2) กำหนดเป้าหมายในการวิเคราะห์ข้อมูล 3) จัดเตรียมข้อมูล 4) วิเคราะห์ข้อมูล 5) การทำโมเดลการทำนายข้อมูล เครื่องมือที่ใช้ในการวิเคราะห์ข้อมูล ได้แก่ Connect X สำหรับรวบรวมข้อมูล Jupyter Notebook และ Microsoft Excel สำหรับการทำความสะอาดข้อมูล และการทำโมเดลทำนาย

1.2 วัตถุประสงค์ของปริญญานิพนธ์

เพื่อวิเคราะห์พฤติกรรมของลูกค้าที่จะนำไปสู่การตัดสินใจซื้อทรัพย์สินและการสร้างแบบจำลองการคาดการณ์

1.3 ขอบเขตของปริญญานิพนธ์

1.3.1 ชุดข้อมูลสำหรับการวิเคราะห์เป็นข้อมูลการเข้าใช้งานเว็บไซต์ของบริษัทสินทรัพย์แห่งหนึ่ง

1.3.2 ขั้นตอนการวิเคราะห์ข้อมูลแบ่งเป็น 4 กระบวนการ ดังนี้

1.3.2.1 Data Understanding

การศึกษาและทำความเข้าใจข้อมูลว่าประกอบด้วยข้อมูลอะไรบ้าง มีชนิดข้อมูลและรูปแบบของข้อมูลเป็นอย่างไร

1.3.2.2 Data Preparation

การจัดเตรียมข้อมูลและทำความสะอาดข้อมูล (Data Cleaning) ให้พร้อมสำหรับการวิเคราะห์ข้อมูล

1.3.2.3 Data Analytics

การวิเคราะห์ข้อมูลพฤติกรรมใดที่มีผลต่อการตัดสินใจซื้อของลูกค้า

1.3.3 สร้างโมเดลทำนายผล

1.3.3.1 คัดเลือกอัลกอริทึมเพื่อใช้ในการทำนายข้อมูล ประกอบด้วย

- Random Forest
- Decision Tree
- Logistic Regression
- Support Vector Machine

1.3.3.2 สร้างโมเดลทำนายผลด้วยอัลกอริทึมที่กำหนด

1.3.3.3 เลือกโมเดลทำนายผลที่ให้ค่าความแม่นยำสูงสุด

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ทำให้ทราบถึงพฤติกรรมที่มีผลทำให้ลูกค้าตัดสินใจซื้อทรัพย์สิน

1.4.2 ทำให้ได้โมเดลสำหรับการพยากรณ์การตัดสินใจซื้อทรัพย์สินของลูกค้า

1.4.3 สามารถนำผลการวิเคราะห์ไปใช้เพื่อการตัดสินใจและวางแผนทางการตลาดได้

1.5 ขั้นตอนและวิธีการดำเนินงานปริญญานิพนธ์

1.5.1 การรวบรวมข้อมูลและศึกษาข้อมูล (Data Understanding)

ทำการศึกษาข้อมูลโดยข้อมูลที่นำมาทำการวิเคราะห์เป็นข้อมูลพฤติกรรมของผู้ใช้ในการเข้าใช้งานเว็บไซต์และการตัดสินใจซื้อทรัพย์สินของบริษัทสินทรัพย์แห่งหนึ่ง เพื่อวิเคราะห์พฤติกรรมและสร้างโมเดลที่ช่วยในการทำนายพฤติกรรมการซื้อในอนาคต โดยทำความเข้าใจข้อมูลแต่ละแอททริบิวต์ถึงความหมาย รูปแบบของข้อมูล และชนิดของข้อมูลที่จัดเก็บ จากนั้นทำการตรวจสอบรูปแบบข้อมูล ระบุปัญหาของข้อมูล แล้วนำเสนอข้อมูลเบื้องต้นและสมมุติฐานในการวิเคราะห์ให้กับอาจารย์ที่ปรึกษา

1.5.2 การเตรียมข้อมูล (Data Preparation)

นำข้อมูลที่ได้ทำการรวบรวมมาในขั้นตอนก่อนหน้านี้ มาทำ Data Cleaning เพื่อให้ข้อมูลมีคุณภาพเหมาะสมในการนำไปทำการวิเคราะห์ เป็นกระบวนการตรวจสอบ จัดการ และแก้ไขรายการข้อมูลที่ไม่ถูกต้อง ข้อมูลว่าง ข้อมูลที่ซ้ำซ้อน หรือข้อมูลที่มีความผิดปกติ ซึ่งเป็นกระบวนการที่สำคัญในกระบวนการวิเคราะห์ข้อมูล ต้องมีการแทนที่ การปรับปรุง หรือการลบข้อมูลที่ไม่ถูกต้องออกไป เพื่อให้ข้อมูลมีคุณภาพและปรับเปลี่ยนโครงสร้างข้อมูลให้เป็นรูปแบบที่เหมาะสมต่อการนำไปวิเคราะห์ โดยในการเตรียมข้อมูลนี้ใช้ภาษา Python และ Jupyter Notebook

1.5.3 การวิเคราะห์ข้อมูล (Data Analytics)

เมื่อทำการเตรียมข้อมูลเรียบร้อยแล้ว นำข้อมูลไปทำการวิเคราะห์เพื่อหาพฤติกรรมที่มีผลต่อการตัดสินใจซื้อทรัพย์สินของลูกค้า โดยการหาความสัมพันธ์ระหว่างข้อมูล โดยใช้ภาษา Python และ Jupyter Notebook

1.5.4 การสร้างโมเดลเพื่อการทำนาย (Data Prediction)

การทำโมเดลการทำนายมีวัตถุประสงค์เพื่อสร้างเครื่องมือที่สามารถคาดการณ์การตัดสินใจซื้อของลูกค้า โดยอิงจากปัจจัยหรือพฤติกรรมที่วิเคราะห์ได้ ขั้นตอนนี้ประกอบด้วยการเลือกโมเดลการทำนายที่เหมาะสม การปรับแต่งพารามิเตอร์ (Hyperparameter Tuning) และการประเมินผลลัพธ์ของโมเดลเพื่อให้ได้ผลลัพธ์ที่แม่นยำที่สุด โดยโมเดลที่นำมาใช้มีทั้งหมด 4 โมเดล ได้แก่

1.5.4.1 Random Forest โมเดลที่ใช้ต้นไม้การตัดสินใจหลายต้น (Decision Trees) เพื่อสร้างผลลัพธ์ที่มีเสถียรภาพและลดความเสี่ยงจาก Overfitting

1.5.4.2 Decision Tree โมเดลที่สร้างต้นไม้เพียงต้นเดียวเพื่อทำการตัดสินใจบนพื้นฐานของกฎที่ชัดเจน

1.5.4.3 Logistic Regression โมเดลเชิงสถิติที่เหมาะสำหรับการพยากรณ์ข้อมูลแบบ Classification โดยเฉพาะการทำนายกลุ่มข้อมูลแบบ Binary

1.5.4.4 Support Vector Machine (SVM) โมเดลที่สร้างเส้นแบ่งระหว่างคลาสข้อมูลโดยใช้ Margin ที่ใหญ่ที่สุดเพื่อเพิ่มความแม่นยำ

1.5.5 จัดทำเอกสาร (Create Document)

เป็นการจัดทำเอกสารเพื่อนำเสนอแนวทางในการจัดทำปริญญานิพนธ์ โดยมีวิธีการและขั้นตอนการดำเนินงาน รวมถึงผลสรุปที่ได้จากการวิเคราะห์ข้อมูลที่สามารถใช้เป็นแหล่งอ้างอิงและทำการวิเคราะห์ข้อมูลต่อไปได้

1.6 ระยะเวลาในการดำเนินงานปริญญานิพนธ์

ตารางที่ 1.1 ขั้นตอนและระยะเวลาในการดำเนินงานปริญญานิพนธ์

ขั้นตอนการดำเนินงาน	ส.ค. 67	ก.ย. 67	ต.ค. 67	พ.ย. 67	ธ.ค. 67
1. การรวบรวมและศึกษาข้อมูล	←	→			
2. การเตรียมข้อมูล		←	→		
3. การวิเคราะห์ข้อมูล			←	→	
4.การสร้างโมเดลเพื่อการทำนาย				←	→
5.จัดทำเอกสารปริญญานิพนธ์					← →

1.7 อุปกรณ์และเครื่องมือที่ใช้ในการวิเคราะห์ข้อมูล

1.7.1 ฮาร์ดแวร์

1.7.1.1 เครื่องคอมพิวเตอร์โน้ตบุค MSI

- Intel Core I7-7
- Ram 12 GB
- SSD 128 GB
- Window 10 Pro

1.7.2 ซอฟต์แวร์

1.7.2.1 ระบบปฏิบัติการ Window 10 Pro

1.7.2.2 โปรแกรม Jupyter Notebook (Python 3.0)

1.7.2.3 โปรแกรม Microsoft Excel

1.7.2.4 Connect X

บทที่ 2
การทบทวนเอกสารงานวิจัยและวรรณกรรมที่เกี่ยวข้อง

ในการจัดทำปริญญานิพนธ์นี้ ผู้จัดทำได้ทำการศึกษาค้นคว้า แนวคิด ทฤษฎี หลักการ และเครื่องมือต่างๆ สำหรับเป็นแนวทางในการจัดทำ โดยประกอบด้วย

2.1 Data Science[1]

วิทยาการข้อมูล (Data Science) เป็นสาขาที่ใช้หลักการทางคณิตศาสตร์ สถิติ และเทคโนโลยีในการวิเคราะห์และสกัดความรู้จากข้อมูลจำนวนมาก โดยมีเป้าหมายเพื่อค้นหารูปแบบแนวโน้ม และข้อมูลเชิงลึกที่สามารถนำไปใช้ตัดสินใจหรือแก้ไขปัญหาได้ วิทยาการข้อมูลผสมผสานศาสตร์หลายด้าน เช่น การวิเคราะห์ข้อมูล (Data Analytics) ปัญญาประดิษฐ์ (Artificial Intelligence) และการเรียนรู้ของเครื่อง (Machine Learning) ทำให้สามารถนำไปประยุกต์ใช้ในหลายอุตสาหกรรม เช่น การตลาด การเงิน การแพทย์ และการผลิต

<u>กระบวนการทำงานของวิทยาการข้อมูล</u>

1. การเก็บรวบรวมข้อมูล (Data Collection) – รวบรวมข้อมูลจากแหล่งต่าง ๆ เช่น ฐานข้อมูล เว็บไซต์ API หรืออุปกรณ์ IoT
2. การทำความเข้าใจข้อมูล (Data Understanding) – วิเคราะห์โครงสร้างของข้อมูล ตรวจสอบค่าที่ขาดหายไป และทำการวิเคราะห์เชิงสถิติ
3. การเตรียมข้อมูล (Data Preparation) – ทำความสะอาดข้อมูล แปลงข้อมูล และเลือกคุณสมบัติที่สำคัญสำหรับการวิเคราะห์
4. การสร้างโมเดล (Modeling) – ใช้เทคนิคทางสถิติหรืออัลกอริทึม Machine Learning ในการสร้างแบบจำลองที่สามารถทำนายหรือจำแนกข้อมูล
5. การประเมินผล (Evaluation) – ทดสอบและปรับปรุงโมเดลเพื่อให้ได้ผลลัพธ์ที่แม่นยำและเชื่อถือได้
6. การนำไปใช้งาน (Deployment) – นำโมเดลที่พัฒนาไปใช้ในระบบจริง พร้อมทั้งติดตามและปรับปรุงให้มีประสิทธิภาพอยู่เสมอ

[1] https://th.wikipedia.org/wiki/วิทยาการข้อมูล

2.2 Data Collection [2]

กระบวนการเก็บรวบรวมข้อมูลจากแหล่งต่าง ๆ เพื่อนำมาใช้ในการวิเคราะห์ ประมวลผล หรือจัดเก็บเพื่อใช้ในการตัดสินใจ การเก็บข้อมูลสามารถทำได้หลายวิธี เช่น การสัมภาษณ์ผู้เกี่ยวข้อง การสำรวจแบบสอบถาม การเก็บข้อมูลจากฐานข้อมูลออนไลน์ หรือการเก็บข้อมูลจากการวิเคราะห์ข้อมูลที่มีอยู่แล้ว การเก็บข้อมูลที่ถูกต้องและเพียงพอเป็นสิ่งสำคัญในการใช้ข้อมูลให้เกิดประโยชน์สูงสุดในการทำงานต่อไป ซึ่งข้อมูลมีด้วยกัน 2 ประเภท

1. Quantitative Data (ข้อมูลเชิงปริมาณ)

ข้อมูลเชิงปริมาณ คือ ข้อมูลที่ได้จากการเก็บข้อมูลออกมาในเชิงตัวเลข (Numerical data) เพื่อแสดงปริมาณของสิ่งที่นับหรือสิ่งที่วัดได้ สามารถแบ่งได้เป็น 2 ประเภท

- ข้อมูลปริมาณแบบต่อเนื่อง (Continuous Data) เป็นข้อมูลที่มีค่าต่อเนื่องกันในช่วงที่กำหนด เช่น อายุ น้ำหนัก ส่วนสูง
- ข้อมูลเชิงปริมาณแบบไม่ต่อเนื่อง (Discrete Data) เป็นข้อมูลจำนวนเต็มหรือจำนวนนับ เช่น จำนวนรถยนต์ในกรุงเทพฯ จำนวนนักศึกษา จำนวนสมาชิกในครอบครัว

2. Qualitative Data (ข้อมูลเชิงคุณภาพ)

ข้อมูลเชิงคุณภาพ คือ ข้อมูลที่ไม่สามารถวัดค่าได้ด้วยตัวเลขว่ามากหรือน้อย แต่จะเป็นข้อมูลที่แสดงถึงสถานภาพ คุณลักษณะ ทัศนคติ หรือคุณสมบัติ มักจะอยู่ในรูปแบบของคำพูด การบรรยาย การอธิบาย ตัวหนังสือ รูปภาพ หรือสัญลักษณ์ต่าง ๆ เช่น สี สถานที่ที่ชอบไป เชื้อชาติ สถานภาพสมรส ศาสนา กลุ่มเลือด

2.3 Data Understanding [3]

กระบวนการทำความเข้าใจข้อมูล (Data Understanding) เป็นขั้นตอนสำคัญในงานวิเคราะห์ข้อมูล โดยเริ่มต้นจากการรวบรวมข้อมูลจากแหล่งต่าง ๆ และทำการสำรวจเบื้องต้นเพื่อทำความเข้าใจโครงสร้างและลักษณะของข้อมูล ขั้นตอนนี้มักใช้การตรวจสอบประเภทของข้อมูล เช่น ข้อมูลเชิงปริมาณ (Numerical Data) หรือข้อมูลเชิงคุณภาพ (Categorical Data) รวมถึงการวิเคราะห์สถิติเบื้องต้น เช่น ค่าเฉลี่ย ค่ามัธยฐาน ส่วนเบี่ยงเบนมาตรฐาน และการกระจายของข้อมูล นอกจากนี้ยังอาจใช้การวิเคราะห์ภาพ (Data Visualization) เช่น การสร้างกราฟฮิสโตแกรม กล่องสรุป (Box Plot) และแผนภูมิกระจาย (Scatter Plot) เพื่อให้เห็นแนวโน้มและความสัมพันธ์ของข้อมูล

---

[2] https://www.dittothailand.com/dittonews/gov-what-is-data-collection/

[3] https://www.coraline.co.th/single-post/data-understanding-process

นอกจากการสำรวจลักษณะของข้อมูลแล้ว ขั้นตอนนี้ยังรวมถึงการตรวจสอบความสมบูรณ์ของข้อมูล (Data Quality) เช่น การหาค่าที่หายไป (Missing Values) การตรวจจับค่าผิดปกติ (Outliers) และการระบุความไม่สอดคล้องกันของข้อมูล ซึ่งจะช่วยให้สามารถปรับปรุงคุณภาพของข้อมูลก่อนนำไปใช้ในการวิเคราะห์ นอกจากนี้ อาจมีการตรวจสอบความสัมพันธ์ระหว่างตัวแปร (Correlation Analysis) เพื่อทำความเข้าใจว่าข้อมูลตัวแปรใดมีผลกระทบต่อกัน และสามารถใช้ในการพัฒนาโมเดลได้อย่างมีประสิทธิภาพ การทำความเข้าใจข้อมูลที่ดีจะช่วยให้การวิเคราะห์ในขั้นตอนถัดไปมีความแม่นยำและเชื่อถือได้มากขึ้น

2.4 Data Preparation [4]

กระบวนการเตรียมข้อมูลในงานวิเคราะห์ข้อมูลเป็นขั้นตอนสำคัญที่ช่วยให้ข้อมูลพร้อมใช้งานสำหรับการวิเคราะห์และการสร้างโมเดล กระบวนการนี้เริ่มต้นด้วยการรวบรวมข้อมูลจากแหล่งต่าง ๆ เช่น ฐานข้อมูล ระบบคลาวด์ ไฟล์ CSV หรือ API จากนั้นต้องทำความสะอาดข้อมูลเพื่อลบค่าที่ผิดพลาด แก้ไขข้อมูลที่ขาดหายไป และจัดรูปแบบข้อมูลให้เหมาะสมกับการใช้งาน นอกจากนี้ อาจต้องทำการแปลงข้อมูล (Data Transformation) เช่น การทำให้ข้อมูลอยู่ในรูปแบบที่สม่ำเสมอ การเข้ารหัสค่าที่เป็นข้อความเป็นตัวเลข และการรวมชุดข้อมูลจากหลายแหล่งเพื่อให้มีโครงสร้างที่สมบูรณ์ขึ้น

หลังจากทำความสะอาดและแปลงข้อมูลเรียบร้อยแล้ว ขั้นตอนถัดไปคือการเลือกคุณลักษณะของข้อมูล (Feature Selection) และการลดมิติของข้อมูล (Dimensionality Reduction) เพื่อลดความซับซ้อนและเพิ่มประสิทธิภาพในการวิเคราะห์ จากนั้นทำการแบ่งชุดข้อมูลเป็นชุดข้อมูลผู้ฝึก (Training Dataset) และชุดข้อมูลทดสอบ (Testing Dataset) หากต้องใช้ในการสร้างโมเดลแมชชีนเลิร์นนิง และขั้นตอนสุดท้ายเป็นการตรวจสอบคุณภาพข้อมูลและความสมบูรณ์ของข้อมูลก่อนนำไปวิเคราะห์หรือสร้างโมเดล เพื่อให้แน่ใจว่าผลลัพธ์ที่ได้มีความแม่นยำและเชื่อถือได้

การเตรียมข้อมูลที่มีประสิทธิภาพ ควรมีลักษณะสำคัญดังนี้

- ให้ผลลัพธ์ที่ครบถ้วนสมบูรณ์
- ให้ความสำคัญกับนิยามข้อมูล
- จดบันทึกขั้นตอนการเตรียมข้อมูลโดยละเอียด
- ปรับกระบวนการให้เป็นอัตโนมัติให้มากที่สุด

[4] https://bzinsight.wordpress.com/2014/06/11/การทำ-data-preparation-อย่างมืออาชีพ/

2.5 Data Analytics [5]

การวิเคราะห์ข้อมูลเพื่อทำนายสิ่งที่คาดว่าจะเกิดขึ้นในอนาคต เป็นประโยชน์ในการพัฒนาทางด้านการตลาดที่ตรงใจลูกค้ามากยิ่งขึ้น การวิเคราะห์ข้อมูลเป็นเครื่องมือสำหรับธุรกิจอัจฉริยะ (Business Intelligence) รูปแบบของการวิเคราะห์ข้อมูล (Data Analytics) สามารถแบ่งได้ดังนี้

การวิเคราะห์ข้อมูลแบบพื้นฐาน (Descriptive analytics) เป็นการวิเคราะห์ เพื่อแสดงผลของรายการทางธุรกิจ เหตุการณ์ หรือกิจกรรมต่างๆ ที่ได้เกิดขึ้น หรืออาจกำลัง เกิดขึ้นในลักษณะที่ง่ายต่อการเข้าใจ หรือต่อการตัดสินใจ ตัวอย่างเช่น รายงานการขาย รายงานผล การดำเนินงาน

การวิเคราะห์แบบเชิงวินิจฉัย (Diagnostic analytics) เป็นการอธิบายถึงสาเหตุของสิ่งที่เกิดขึ้น ปัจจัยต่างๆ และความสัมพันธ์ของปัจจัยหรือตัวแปรต่างๆ ที่มีความสัมพันธ์ต่อกันของสิ่งที่เกิดขึ้น ตัวอย่างเช่น ความสัมพันธ์ระหว่างยอดขายต่อกิจกรรมทางการตลาดแต่ละประเภท ซึ่งเป็นก้าวใหม่ที่ช่วยเสริมให้ตัดสินใจไปในทางที่ถูกต้อง

การวิเคราะห์แบบพยากรณ์ (Predictive analytics) เป็นการวิเคราะห์เพื่อพยากรณ์สิ่งที่กำลังจะเกิดขึ้นหรือน่าจะเกิดขึ้น โดยใช้ข้อมูลที่ได้เกิดขึ้นแล้วกับแบบจำลองทางสถิติ หรือ เทคโนโลยีปัญญาประดิษฐ์ต่างๆ (Artificial intelligence) ตัวอย่างเช่น การพยากรณ์ยอดขาย การพยากรณ์ผลประชามติ

การวิเคราะห์แบบให้คำแนะนำ (Prescriptive analytics) เป็นการวิเคราะห์ข้อมูลที่มีความซับซ้อนที่สุด เป็นทั้งการพยากรณ์สิ่งต่างๆ ที่จะเกิดขึ้น ข้อดี ข้อเสีย สาเหตุ และระยะเวลาของสิ่งที่จะเกิดขึ้น ร่วมถึงการให้คำแนะนำทางเลือกต่างๆ ที่มีอยู่ และผลของแต่ละทางเลือก

2.6 Python Language[6]

เป็นภาษาคอมพิวเตอร์ที่นิยมใช้ในการพัฒนาซอฟต์แวร์ (Software Development) วิทยาการข้อมูล (Data Science) และแมชชีนเลิร์นนิง (Machine Learning) นักพัฒนาใช้ Python เนื่องจากเป็นภาษาที่มีประสิทธิภาพ เรียนรู้ง่าย และสามารถทำงานบนแพลตฟอร์มต่างๆ ได้หลากหลาย

ภาษา Python ในการพัฒนาแอปพลิเคชัน ดังตัวอย่างเช่น

2.6.1 การพัฒนาเว็บฝั่งเซิร์ฟเวอร์

---

[5] https://affinity.co.th/data-analytics/

[6] https://aws.amazon.com/th/what-is/python/

การพัฒนาเว็บฝั่งเซิร์ฟเวอร์ประกอบด้วยฟังก์ชัน Backend ที่ซับซ้อน ซึ่งเว็บไซต์ดำเนินการเพื่อแสดงข้อมูลต่อผู้ใช้ ตัวอย่างเช่น เว็บไซต์ต้องโต้ตอบกับฐานข้อมูล สื่อสารกับเว็บไซต์อื่น และปกป้องข้อมูลเมื่อส่งข้อมูลผ่านเครือข่าย

Python มีประโยชน์สำหรับการเขียนชุดคำสั่งฝั่งเซิร์ฟเวอร์ เนื่องจากมีไลบรารีจำนวนมากที่ประกอบด้วยชุดคำสั่งที่เขียนไว้ล่วงหน้าสำหรับฟังก์ชัน Backend ที่ซับซ้อน นักพัฒนาใช้เฟรมเวิร์ก Python ที่หลากหลายซึ่งมีเครื่องมือที่จำเป็นทั้งหมดเพื่อสร้างเว็บแอปพลิเคชันได้เร็วขึ้นและง่ายขึ้นอีกด้วย ตัวอย่างเช่น นักพัฒนาสามารถสร้างโครงสร้างเว็บแอปพลิเคชันได้ภายในไม่กี่วินาที เนื่องจากไม่จำเป็นต้องเขียนขึ้นใหม่ทั้งหมด จากนั้นนักพัฒนาสามารถทดสอบได้โดยใช้เครื่องมือทดสอบของเฟรมเวิร์ก โดยไม่ต้องพึ่งพาเครื่องมือทดสอบภายนอก

2.6.2 ระบบอัตโนมัติด้วยสคริปต์ Python

ภาษาการเขียนสคริปต์คือภาษาการเขียนโปรแกรมที่ทำให้งานที่มนุษย์ทำตามปกติเป็นไปโดยอัตโนมัติ นักพัฒนาจึงใช้สคริปต์ Python อย่างแพร่หลายในการทำงานประจำวัน เช่น

1. การเปลี่ยนชื่อไฟล์จำนวนมากพร้อมกัน
2. การแปลงไฟล์เป็นไฟล์ประเภทอื่น
3. การลบคำที่ซ้ำกันในไฟล์ข้อความ
4. การดำเนินการทางคณิตศาสตร์ขั้นพื้นฐาน
5. การส่งข้อความอีเมล
6. การดาวน์โหลดเนื้อหา
7. การดำเนินการวิเคราะห์บันทึกพื้นฐาน
8. การค้นหาข้อผิดพลาดในหลายไฟล์

2.6.3 วิทยาการข้อมูลและแมชชีนเลิร์นนิง

วิทยาการข้อมูลดึงความรู้อันมีคุณค่าจากข้อมูล และแมชชีนเลิร์นนิง (ML) จะสอนคอมพิวเตอร์ให้เรียนรู้จากข้อมูลโดยอัตโนมัติและทำนายได้อย่างแม่นยำ นักวิทยาศาสตร์ข้อมูลใช้ Python สำหรับงานด้านวิทยาศาสตร์ข้อมูลต่างๆ เช่น

1. การแก้ไขและลบข้อมูลที่ไม่ถูกต้อง ซึ่งเรียกว่าการทำความสะอาดข้อมูล
2. การแยกและเลือกคุณสมบัติต่างๆ ของข้อมูล
3. การระบุประเภทข้อมูล ซึ่งเป็นการเพิ่มชื่อที่มีความหมายสำหรับข้อมูล

4. การค้นหาสถิติต่างๆ จากข้อมูล
5. การแสดงข้อมูลด้วยภาพโดยใช้แผนภูมิและกราฟ เช่น แผนภูมิเส้น กราฟแท่ง ฮิสโทแกรม และแผนภูมิวงกลม

นักวิทยาการข้อมูล (Data Scientist) ใช้ไลบรารี Python ML เพื่อฝึกฝนโมเดล ML และสร้างตัวจำแนกที่จำแนกประเภทข้อมูลได้อย่างแม่นยำ บุคคลในแวดวงต่างๆ ใช้ตัวจำแนกแบบ Python เพื่อทำงานด้านการจำแนกประเภท เช่น การจำแนกประเภทรูปภาพ ข้อความ และการรับส่งข้อมูลทางเครือข่าย การรู้จำเสียง และการจดจำใบหน้า นักวิทยาศาสตร์ข้อมูลยังใช้ Python สำหรับ ดีปเลิร์นนิง ซึ่งเป็นเทคนิค ML ขั้นสูง

2.7 Data Correlation[7]

การวิเคราะห์ข้อมูลนั้นมิใช่เรื่องใหม่หรือเป็นปรากฏการณ์ใหม่ที่เพิ่งเกิดขึ้นในยุคดิจิทัล เพียงแต่ปัจจุบันมีข้อมูลจำนวนมหาศาล ดังนั้นการวิเคราะห์ข้อมูลในปัจจุบันของหลายหน่วยงาน ไม่ว่าจะเป็นหน่วยงานรัฐหรือหน่วยงานเอกชนได้กลายมาเป็นการวิเคราะห์ข้อมูลจากข้อมูลที่มีขนาดใหญ่ (Big Data) วิธีทางสถิติที่นำมาใช้วิเคราะห์ความสัมพันธ์กับข้อมูลขนาดใหญ่ (Correlation Analysis in Big Data) ได้แก่

- การหาค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation) ประเภทต่าง ๆ
- การหาค่า Maximal Information Coefficient (MIC)

ค่า Pearson Correlation เป็นวิธีทางสถิติที่นิยมนำมาใช้วิเคราะห์เพื่อหาความสัมพันธ์ของข้อมูลมากที่สุดวิธีหนึ่ง เนื่องจากเป็นวิธีที่เข้าใจง่ายและสามารถคำนวนได้ไม่ยาก โดยค่า Pearson Correlation จะมีค่าอยู่ระหว่าง -1.0 ถึง +1.0 ซึ่งหากมีค่าใกล้ -1.0 นั้นหมายความว่าตัวแปรทั้งสองตัวมีความสัมพันธ์กันอย่างมากในเชิงตรงกันข้าม หากมีค่าใกล้ +1.0 นั้นหมายความว่า ตัวแปรทั้งสองมีความสัมพันธ์กันอย่างมากในทิศทางเดียวกัน และหากมีค่าเป็น 0 นั้นหมายความว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์ต่อกัน ทั้งนี้ค่าสัมประสิทธิ์สหสัมพันธ์เป็นการวิเคราะห์ความสัมพันธ์ของข้อมูลในลักษณะแบบเส้นตรงเท่านั้น ดังนั้นค่า Pearson Correlation จึงขาดความน่าเชื่อถือในกรณีที่ข้อมูลที่นำมาวิเคราะห์นั้นมิได้มีลักษณะความสัมพันธ์แบบเส้นตรง

ค่า Spearman Correlation เป็นการหาความสัมพันธ์ระหว่างตัวแปร 2 ตัวที่อยู่ในมาตราการวัดระดับ Ordinal Scale โดยค่า Spearman Correlation จะมีค่าอยู่ระหว่าง -1.0 ถึง +1.0 เช่นเดียวกับค่า Pearson Correlation ส่วนการแปลความหมายก็ไม่ต่างจากการแปลความหมายของค่า Pearson Correlation นั้นคือ หากค่า Spearman Correlation ใกล้ -1.0 นั้นหมายความว่าตัวแปรทั้งสองตัวมี

[7] https://bdi.or.th/big-data-101/correlation-analysis-in-big-data/

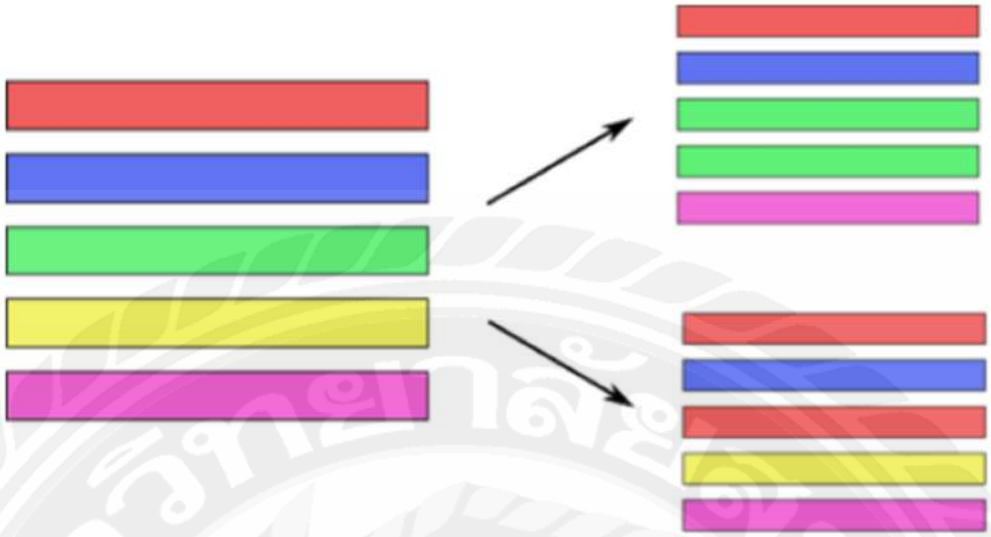
ความสัมพันธ์กันอย่างมากในเชิงตรงกันข้าม หากค่า Spearman Correlation ใกล้ +1.0 นั้นหมายความว่า ตัวแปรทั้งสองมีความสัมพันธ์กันอย่างมากในทิศทางเดียวกัน และหากค่า Spearman Correlation เป็น 0 นั้นหมายความว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์ต่อกัน

ค่า Kendall Correlation เป็นการหาความสัมพันธ์ระหว่างตัวแปร 2 ตัวที่อยู่ในมาตราการวัดระดับ Ordinal Scale โดยค่า Kendall Correlation จะมีค่าอยู่ระหว่าง -1.0 ถึง +1.0 เช่นเดียวกับ ค่า Pearson Correlation และ ค่า Spearman Correlation ส่วนการแปลความหมายก็ไม่ต่างจากการแปลความหมายของค่า Pearson Correlation และ ค่า Spearman Correlation นั้นคือ หากค่า Kendall Correlation ใกล้ -1.0 นั้นหมายความว่าตัวแปรทั้งสองตัวมีความสัมพันธ์กันอย่างมากในเชิงตรงกันข้าม หากค่า Kendall Correlation ใกล้ +1.0 นั้นหมายความว่า ตัวแปรทั้งสองมีความสัมพันธ์กันอย่างมากในทิศทางเดียวกัน และหากค่า Kendall Correlation เป็น 0 นั้นหมายความว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์ต่อกัน อย่างไรก็ตาม ค่า Kendall Correlation จะสามารถใช้ในการบ่งบอกถึงระดับความเข้มข้น (strength) ของความสัมพันธ์ระหว่างตัวแปรทั้งสองได้ดีกว่าค่า Pearson Correlation และ ค่า Spearman Correlation ดังนั้นโดยปกติค่า Kendall Correlation จะน้อยกว่า ค่า Pearson Correlation และ ค่า Spearman Correlation เสมอ

2.8 Random Forest Classifier[8]

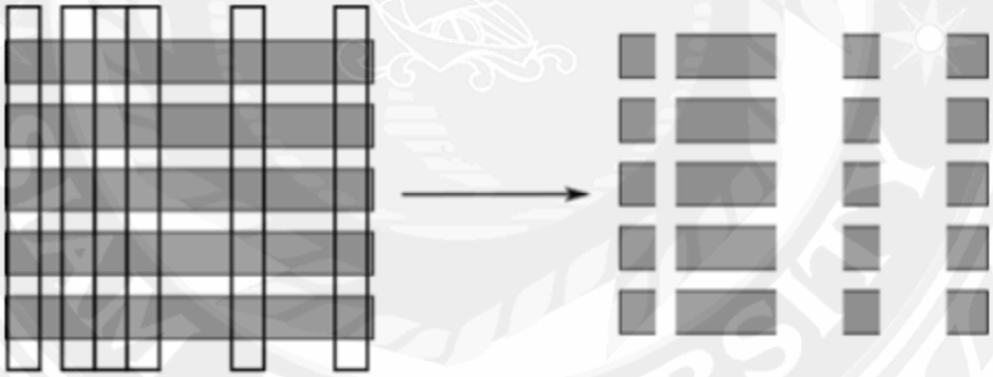
เป็นแบบจำลอง (Model) ประเภทหนึ่งของ Machine Learning ถูกพัฒนาขึ้นจาก Decision Tree โดยมีความต่างกันตรงที่ Random Forest เป็นการเพิ่มจำนวน Tree เป็น Tree หลายๆ ต้น ทำให้ประสิทธิภาพในการทำงานสูงขึ้น แม่นยำมากขึ้น ซึ่งโมเดล Random Forest เป็นโมเดลที่ได้รับความนิยมไปอย่างมากในการใช้ Machine Learning

[8] https://medium.com/@pradyasin/random-forest-คืออะไร-74d2a0af3d7



รูปที่ 2.1 ตัวอย่างการแบ่งข้อมูลออกเป็น Tree แต่ละต้น

คล้ายกับ Bagging โดยBagging จะมีการแบ่งข้อมูลออกเป็น Tree หลายๆ ต้น แต่การทำ Bagging จะมีปัญหาเรื่องความไม่เป็นอิสระของข้อมูลเนื่องจากต่อให้ แยกออกไปหลายๆ Tree ก็จริงแต่มันก็คือข้อมูลเดียวกัน Random Forest จึงเข้ามาแก้ปัญหาตรงนี้ โดยการทำ Random Sample Feature โดย



รูปที่ 2.2 ตัวอย่างการทำ Random Sample Feature

นอกจากจะแบ่งเป็น Tree หลายๆ ต้นแล้ว ยังแบ่ง Feature ของ Tree แต่ละต้นจะมี Feature ที่ไม่เหมือนกันทั้งหมด เพื่อทำให้แต่ละ Tree มีความหลากหลายและมีความอิสระกันมากขึ้น

2.9 Decision Tree Classifier[9]

ต้นไม้ตัดสินใจ คือ การจำลองวิธีการตัดสินใจของมนุษย์ ซึ่งการตัดสินใจแต่ละครั้งของมนุษย์ มนุษย์จะแตกโจทย์หลักออกเป็นโจทย์ย่อยหลายๆ โจทย์ก่อนเพื่อที่จะได้ง่ายต่อการตัดสินใจ หรือจะนำเอาปัจจัยต่างๆ ที่เกี่ยวข้องกับการตัดสินใจหรือเกี่ยวข้องกับโจทย์หลักมาตั้งเป็นคำถามใหม่หรือแตกเป็นโจทย์ย่อย และถามตัวเองใหม่อีกครั้ง เช่น วันนี้, มีเพื่อนมาถาม ว่าไปกินข้าวมันไก่ด้วยกันไหม? มนุษย์จะคิดก่อนว่า อร่อยหรือเปล่านะ? ถ้าไม่อร่อยแล้วราคาแพงไหม? แล้วค่อยตัดสินใจว่าจะไปหรือไม่ไป

รูปที่ 2.3 ตัวอย่าง Decision Tree

แบบจำลองต้นไม้ (Model Decision Tree) เป็น Rule-Based Model ที่สร้างเงื่อนไข If-else ขึ้นมาจากข้อมูลในตัวแปร เพื่อที่จะแบ่งข้อมูลออกเป็นกลุ่มใหม่ที่สามารถอธิบาย Target ได้ดีที่สุด โดยการสร้างเงื่อนไข If-else ในแต่ละตัวแปร จะถูกกำหนดด้วย Objective Function ซึ่ง Model Decision Tree มี Objective Function อยู่หลายตัว ตามประเภทของ Decision Tree นั้น ๆ

Decision Tree จะแบ่งออกเป็น 2 ประเภท คือ

1. Regression Tree คือ Decision Tree ที่ใช้สำหรับการทำโจทย์ Regression โดยมีค่า Residual sum of squares (RSS) เป็น Objective Function ในการหาจุดที่ดีที่สุดในการแบ่งข้อมูล (Split point) จากการ Minimize ให้ RSS มีค่าน้อยที่สุด
   - Residual (e_i) คือ ค่าความคลาดเคลื่อน หรือค่า Error ระหว่าง y ทุก ๆ จุดในข้อมูลกับ y_hat ที่ได้มาจากการประมาณค่าขึ้นมาการคำนวณ Residual ของข้อมูลตัวที่ i

[9] https://www.borntodev.com/2022/09/15/รู้จักกับ-decision-tree/

$$e_i = y_i - \hat{y}_i$$

รูปที่ 2.4 สูตรหาค่าความคลาดเคลื่อน

- Residual sum of squares (RSS) คือ การวัดค่า Residual หรือ ค่า Error ของทุกๆ จุดในชุดข้อมูล และนำมายกกำลังสอง เพื่อให้ค่า Residual เป็นบวก และเป็นการทำ Normalize ด้วย เนื่องจากถ้า y_hat มีค่ามากกว่า yi จะทำให้ค่า Residual ติดลบ

$$RSS = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2$$

รูปที่ 2.5 สูตรการคำนวณ Residual sum of squares

2. Classification Tree คือ Decision Tree ที่ใช้สำหรับการทำ Classification โดยจะใช้ Gini Impurity หรือ Entropy เป็น Objective Function ในการหาจุดที่ดีที่สุดในการแบ่งข้อมูล (Split point)
   - Gini Impurity คือ การวัดค่า Impurity หรือ ค่าความไม่บริสุทธิ์ในการอธิบาย Target ของกลุ่มที่ถูกแบ่งออกมาจากตัวแปร นั่นหมายความว่าถ้าค่า Impurity ยิ่งต่ำก็ยิ่งแบ่งข้อมูลออกมาได้ดีนั่นเอง
   - การคำนวณ Gini Impurity คือ การนำเอาผลรวมของค่าความน่าจะเป็นของเหตุการณ์ที่ สนใจมาคูณด้วย (1 ลบ ค่าความน่าจะเป็นของเหตุการณ์ที่ สนใจ)

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

รูปที่ 2.6 การคำนวณ Gini Impurity

หลังจากได้ค่า Gini Impurity ของแต่ละกลุ่มในทุก ๆ ตัวแปรแล้ว จะทำการหาค่า Weighted Gini Impurity เพื่อเลือกตัวแปรที่มีค่า Weighted Gini Impurity ต่ำที่สุดมาใช้ในการตัดสินใจก่อน เพราะสามารถ Split ข้อมูลได้ดีที่สุด

การคำนวณ Weighted Gini Impurity คือ การนำเอาผลรวมของค่า Gini ในเหตุการณ์ที่สนใจกับจำนวนข้อมูลของ Class ที่ i ในตัวแปรที่ สนใจ และหารด้วยจำนวนข้อมูลทั้งหมดในตัวแปรที่สนใจ

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

รูปที่ 2.7 การคำนวณ Weighted Gini Impurity

2.10 Logistic Regression Model[10]

คือเทคนิคการวิเคราะห์ข้อมูลที่ใช้วิชาคณิตศาสตร์เพื่อหาความสัมพันธ์ระหว่างสองปัจจัยข้อมูล จากนั้นจะใช้ความสัมพันธ์นี้เพื่อคาดการณ์ค่าของปัจจัยเหล่านั้นโดยอาศัยปัจจัยอื่นๆ การคาดการณ์มักจะมีจำนวนผลลัพธ์ที่จำกัด เช่น ใช่หรือไม่ เป็นต้น

ตัวอย่างเช่น สมมติว่าคุณต้องการที่จะคาดการณ์ว่าผู้เข้าชมเว็บไซต์ของคุณจะคลิกปุ่มชำระเงินในรถเข็นช้อปปิ้งของพวกเขาหรือไม่ การวิเคราะห์รีเกรสชันโลจิสติกจะพิจารณาพฤติกรรมของผู้เข้าชมในอดีต เช่น เวลาที่ใช้บนเว็บไซต์และจำนวนสินค้าในรถเข็น เป็นต้น โดยกำหนดว่า ในอดีตหากผู้เข้าชมใช้เวลามากกว่า 5 นาทีบนเว็บไซต์และเพิ่มสินค้าลงในรถเข็นมากกว่า 3 รายการ จากนั้นพวกเขาคลิกปุ่มชำระเงิน ด้วยการใช้ข้อมูลนี้ ฟังก์ชันรีเกรสชันโลจิสติกสามารถคาดการณ์พฤติกรรมของผู้เข้าชมเว็บไซต์รายใหม่ได้

รีเกรสชันโลจิสติกเป็นเทคนิคที่สำคัญในสาขาปัญญาประดิษฐ์และแมชชีนเลิร์นนิ่ง (AI/ML) โมเดล ML เป็นโปรแกรมซอฟต์แวร์ที่คุณสามารถฝึกเพื่อดำเนินการประมวลผลข้อมูลที่ซับซ้อนได้โดยไม่มีการแทรกแซงของมนุษย์ โมเดล ML ที่สร้างขึ้นโดยใช้รีเกรสชันโลจิสติกช่วยให้องค์กรได้รับข้อมูลเชิงลึกที่ได้จากข้อมูลในการทำงานของพวกเขา โดยพวกเขาสามารถใช้ข้อมูลเชิงลึกเหล่านี้สำหรับการวิเคราะห์เชิงคาดการณ์เพื่อลดต้นทุนการดำเนินงาน เพิ่มประสิทธิภาพ และปรับขนาดได้เร็วขึ้น ยกตัวอย่างเช่น ธุรกิจจะสามารถพบรูปแบบที่ช่วยปรับปรุงการรักษาพนักงานเดิมหรือนำไปสู่การออกแบบผลิตภัณฑ์ที่มีกำไรมากขึ้น

[10] https://aws.amazon.com/th/what-is/logistic-regression/

ประโยชน์ของการใช้รีเกรสชันโลจิสติกที่ดี

ความเรียบง่าย

โมเดลรีเกรสชันโลจิสติกเป็นคณิตศาสตร์ที่ซับซ้อนน้อยกว่าวิธี ML อื่นๆ ดังนั้นคุณสามารถใช้เทคนิคนี้ได้แม้ไม่มีใครในทีมของคุณมีความเชี่ยวชาญเกี่ยวกับ ML ในเชิงลึกก็ตาม

ความเร็ว

โมเดลรีเกรสชันโลจิสติกสามารถประมวลผลข้อมูลปริมาณมากด้วยความเร็วสูงได้เพราะต้องการความสามารถในการคำนวณน้อยกว่า เช่น หน่วยความจำและกำลังประมวลผล ซึ่งทำให้เหมาะสำหรับองค์กรที่เริ่มต้นโครงการ ML เพื่อความสำเร็จอย่างรวดเร็ว

ความยืดหยุ่น

คุณสามารถใช้รีเกรสชันโลจิสติกเพื่อหาคำตอบสำหรับคำถามที่มีคำตอบสองข้อหรือมากกว่าได้ นอกจากนี้คุณยังสามารถใช้เพื่อประมวลผลข้อมูลล่วงหน้าได้อีกด้วย ตัวอย่างเช่น คุณสามารถเรียงลำดับข้อมูลที่มีช่วงค่ากว้าง เช่น การทำธุรกรรมธนาคารมาเป็นช่วงค่าที่แคบและจำกัดลงด้วยรีเกรสชันโลจิสติกได้ จากนั้น คุณสามารถประมวลผลชุดข้อมูลนี้ที่มีขนาดเล็กโดยใช้เทคนิค ML อื่นๆ เพื่อการวิเคราะห์ที่ถูกต้องมากขึ้น

การแสดงผล

รีเกรสชันโลจิสติกทำให้นักพัฒนาสามารถมองเห็นกระบวนการซอฟต์แวร์ภายในได้มากขึ้นกว่าเทคนิคการวิเคราะห์ข้อมูลอื่นๆ การแก้ไขปัญหาและการแก้ไขข้อผิดพลาดยังง่ายขึ้นเนื่องจากการคำนวณมีความซับซ้อนน้อยกว่า

รีเกรสชันโลจิสติกเป็นหนึ่งในเทคนิคการวิเคราะห์รีเกรสชันที่แตกต่างกันหลายอย่างที่นักวิทยาศาสตร์ข้อมูลนิยมใช้ในแมชชีนเลิร์นนิง (ML) เพื่อให้เข้าใจถึงรีเกรสชันโลจิสติก ก่อนอื่นต้องเข้าใจการวิเคราะห์รีเกรสชันขั้นพื้นฐาน ด้านล่างนี้ จะใช้ตัวอย่างของการวิเคราะห์รีเกรสชันเชิงเส้นเพื่อแสดงให้เห็นถึงวิธีการวิเคราะห์รีเกรสชัน

ระบุคำถาม

การวิเคราะห์ข้อมูลใดๆ เริ่มต้นด้วยคำถามทางธุรกิจ สำหรับรีเกรสชันโลจิสติกคุณควรตีกรอบคำถามเพื่อให้ได้มาซึ่งผลลัพธ์เฉพาะ

- วันที่ฝนตกส่งผลกระทบต่อยอดขายรายเดือนของ หรือไม่ (ใช่หรือไม่)
- ประเภทของกิจกรรมบัตรเครดิตใดที่ลูกค้ากำลังดำเนิน (ถูกต้อง ฉ้อโกง หรืออาจฉ้อโกง)

การเก็บรวบรวมข้อมูลประวัติ

หลังจากระบุคำถามแล้วคุณต้องระบุปัจจัยข้อมูลที่เกี่ยวข้อง จากนั้นคุณจะเก็บรวบรวมข้อมูลที่ผ่านมาของทุกปัจจัย ตัวอย่างเช่น ในการตอบคำถามแรกที่แสดงข้างต้น คุณควรรวบรวมจำนวนวันที่ฝนตกและข้อมูลการขายรายเดือนของคุณแต่ละเดือนในช่วง 3 ปีที่ผ่านมา

ฝึกโมเดลวิเคราะห์รีเกรสชัน

คุณจะประมวลผลข้อมูลประวัติโดยใช้ซอฟต์แวร์รีเกรสชัน ซอฟแวร์จะประมวลผลจุดข้อมูลต่างๆ และเชื่อมต่อทางคณิตศาสตร์โดยใช้สมการ ยกตัวอย่างเช่น ถ้าจำนวนวันที่ฝนตกเป็นเวลาสามเดือนคือ 3, 5, และ 8 และจำนวนยอดขายในเดือนนั้นคือ 8, 12, และ 18 อัลกอริธึมรีเกรสชันจะเชื่อมโยงปัจจัยดังกล่าวเป็นสมการดังนี้

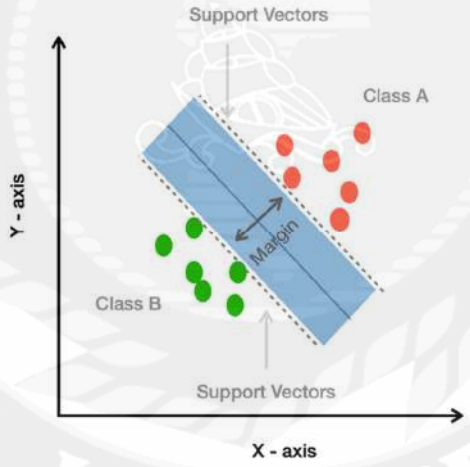
จำนวนการขาย = 2* (จำนวนวันที่ฝนตก) + 2

ทำการคาดคะเนสำหรับค่าที่ไม่รู้จัก

สำหรับค่าที่ไม่รู้จัก ซอฟต์แวร์จะใช้สมการในการคาดคะแน ถ้าคุณรู้ว่าฝนจะตกเป็นเวลาหกวันในเดือนกรกฎาคม, ซอฟแวร์จะประมาณการมูลค่าการขายเดือนกรกฎาคมเป็น 14

2.11 Support Vector Machine[11]

SVM ย่อจาก Support Vector Machine เป็น Machine Learning Algorithm ประเภท Supervised Learning มีเป้าหมาย คือ หา Hyperplane ใน N-dimensional Space โดยที่ N คือ จำนวน Features เพื่อใช้ในการ Classify Data Points



รูปที่ 2.8 ตัวอย่างกรแบ่ง SVM

ข้อดีของ **SVM**

1. มีประสิทธิภาพใน High-dimensional Space
2. ยังคงมีประสิทธิภาพ เมื่อจำนวนของ Dimensions มากกว่า จำนวนของ Sample

[11] https://www.nerd-data.com/svm/

3. ใช้ Subset ของ Training Points (Support Vectors) ทำให้ใช้ Memory ได้อย่างมีประสิทธิภาพ
4. Kernel Functions ที่แตกต่างกัน สามารถใช้ในการกำหนด Decision Function ได้

ข้อดีของ SVM

- จัดการกับ Non-Linear ได้ โดยใช้เทคนิคของ Kernel
- การ Maximized Margin ทำให้เกิดความทนทานที่ดี (Robustness)
- สามารถควบคุม Overfitting โดยใช้เทคนิค Soft Margins

ข้อเสียของ SVM

- ไม่เหมาะกับ Dataset ขนาดใหญ่ เนื่องจากใช้เวลาใน Train ที่นาน
- ประสิทธิภาพจะขึ้นกับการเลือก Kernel
- ต้องมีการทำ Feature Scaling ก่อน เพื่อให้ได้ Accuracy ที่ดี

Hyper-parameters ใน SVM

- **C** หรือ Regularization parameter กรณีค่า C น้อย จะหมายถึง การมี Margin ที่กว้าง ซึ่งอาจส่งผลให้มีการละเมิดเข้ามาใน Margin มากขึ้น กรณีค่า C สูง จะหมายถึง การมี Margin ที่แคบ มีเป้าหมายใน Classify Training Data ให้ถูกมากที่สุด โดยให้ Model มีอิสระในการเลือก Samples เข้ามามากกว่า ในการสร้าง Support Vectors
- **Kernel** คือ ประเภทของ Kernel เช่น ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’
- **Degree** คือ Degree ของ Polynomial Kernel Function (‘poly’) ในกรณี Kernels ประเภทอื่นๆ ไม่ต้องพิจารณาค่านี้
- **Gamma** คือ ค่า Kernel Coefficient ค่า Default คือ ‘scale’ ซึ่งจะถูกคำนวณจาก Data Inputs หากเลือก ‘auto’ จะใช้ 1/n_features

บทที่ 3
การวิเคราะห์ข้อมูล

3.1 รายละเอียดของปริญญานิพนธ์

ในการวิเคราะห์พฤติกรรมของลูกค้าที่มีผลต่อการตัดสินใจซื้อทรัพย์สิน และการสร้างแบบจำลองเพื่อการทำนายจากข้อมูลการใช้งานเว็บของบริษัทสินทรัพย์แห่งหนึ่ง มีวัตถุประสงค์เพื่อทำความเข้าใจพฤติกรรมของลูกค้า รวมถึงปัจจัยที่มีอิทธิพลต่อการตัดสินใจซื้อทรัพย์สิน โดยn ใช้กระบวนการวิเคราะห์ข้อมูลที่เป็นระบบและการนำเสนอผลลัพธ์ในเชิงวิชาการ

หลังจากที่ได้รวบรวมข้อมูลจากแหล่งข้อมูลต่างๆ แล้ว จะทำการตรวจสอบคุณภาพของข้อมูลว่าพร้อมที่จะนำไปทำการวิเคราะห์ต่อไปหรือไม่ โดยได้ทำการหาค่าที่ขาดหายไป (Missing Value) ค่าที่ผิดปกติ (Outlier) ข้อมูลที่ซ้ำซ้อน (Duplicate) ถ้าพบข้อผิดพลาดดังกล่าวจะทำการปรับปรุงแก้ไขข้อมูลให้มีความถูกต้องครบถ้วนให้ได้มากที่สุด เพื่อให้การวิเคราะห์มีความแม่นยำสูง ในการการหาความสัมพันธ์ของข้อมูล (Correlation) และการสร้างแบบจำลองเพื่อการทำนาย (Predictive Model) โดยใช้ Jupyter Notebook และภาษา Python ในการวิเคราะห์ข้อมูลและหาโมเดลเพื่อการทำนาย โดยนำเสนอผลลัพธ์เป็น Dashboard เพื่อให้ง่ายต่อการทำความเข้าใจ

3.2 ขั้นตอนในการวิเคราะห์ข้อมูล

ขั้นตอนในการวิเคราะห์พฤติกรรมของลูกค้าที่มีผลต่อการตัดสินใจซื้อ และการสร้างแบบจำลองการทำนาย มีดังนี้

3.2.1 กำหนดเป้าหมายในการวิเคราะห์ข้อมูล (Problem Definition)

เพื่อให้ทราบถึงพฤติกรรมของลูกค้าที่มีผลต่อการตัดสินใจซื้อสินค้าและบริการ รวมถึงระบุว่าพฤติกรรมใดบ้างที่มีความสัมพันธ์หรือส่งผลโดยตรงต่อการตัดสินใจซื้อ และโมเดลใดเหมาะสมที่สุดสำหรับการวิเคราะห์ข้อมูลชุดนี้เพื่อให้ได้ผลลัพธ์ที่มีความแม่นยำและประสิทธิภาพสูงสุด

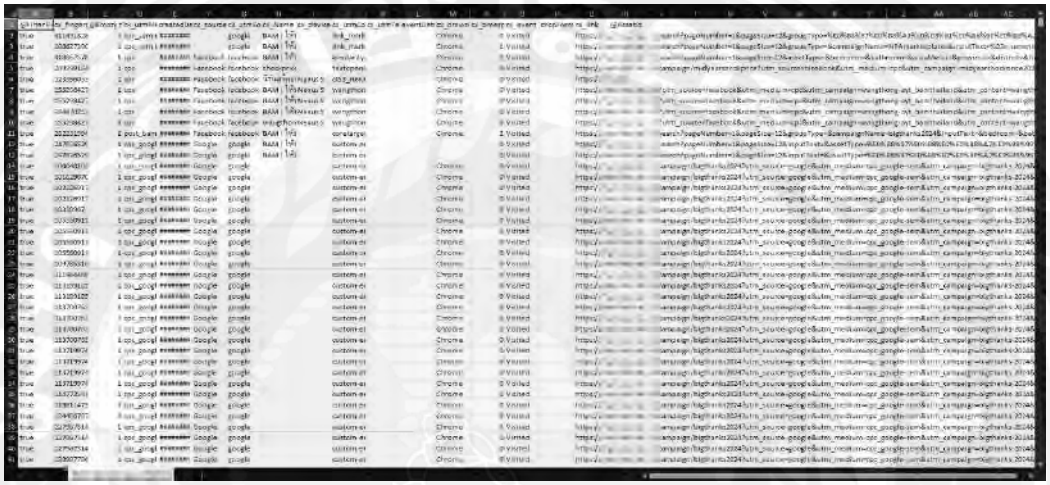
3.2.2 ทำความเข้าใจข้อมูล (Data Understanding)

ผู้จัดทำได้ศึกษาข้อมูลพื้นฐานของชุดข้อมูลพฤติกรรมการใช้งานเว็บไซต์ของบริษัทสินทรัพย์แห่งหนึ่ง และการตัดสินใจซื้อทรัพย์สินของลูกค้า ข้อมูลมีขนาด 18 คอลัมน์ 8,066,118 เรคคอร์ด ซึ่งข้องมูลชุดนี้เป็นข้อมูลของเดือน มกราคม – กรกฎาคม 2024 โดยดึงข้อมูลมากจาก

Connect X ซึ่งคล้ายๆ กับ Database ของบริษัท เมื่อรวบรวมข้อมูลได้ตามที่ต้องการแล้ว จะนำไปสู่ขั้นตอนการเตรียมข้อมูลต่อไป

3.2.3 การเตรียมข้อมูล (Data Preparation)

จากชุดข้อมูลที่ได้รับมานั้นเป็นข้อมูลที่ยังไม่พร้อมที่จะนำไปทำการวิเคราะห์ ผู้จัดทำจึงได้ทำการจัดการและปรับรูปแบบของข้อมูลให้เหมาะสมเพื่อให้ได้ผลการวิเคราะห์ที่แม่นยำที่สุด ในขั้นตอนการเตรียมข้อมูบนี้ผู้จัดทำได้ใช้ไลบรารี่ของ Python ช่วย



รูปที่ 3.1 ตัวอย่างชุดข้อมูล

จากรูปที่ 3.1 ในชุดข้อมูลมีที่ขาดหายไป (Missing Value) เป็นจำนวนมาก จึงได้ทำการพิจารณาทีละรายการว่าควรจะลบหรือเก็บข้อมูลรายการนั้นไว้ เพื่อให้ข้อมูลมีความสมดุลมากที่สุด (Data Balancing) รวมถึงหาค่าที่ผิดปกติ (Outlier) และทำการปรับปรุงแก้ไข หลังจากนั้นทำการคัดเลือกข้อมูลที่ต้องการใช้ เพื่อนำไปเป็น Data ใหม่ ที่เหมาะสมต่อการนำไปวิเคราะห์พฤติกรรมที่มีผลต่อการตัดสินใจซื้อทรัพย์สิน

```
df = df[df['cx_fingerprint'].notna()]
df
```

รูปที่ 3.2 ตัวอย่างชุดคำสั่งสำหรับแสดงแถวที่มีค่าว่างในคอลัมน์ cx_fingerprint

```
df = df.drop(columns=['@Filter Period','@Empty fingerprint','cx_utmMedium','createdDate','cx_source','cx_utmSource','cx_Name',
                      'cx_utmTerm','eventDetail','cx_browserName','cx_timespent','dropFormDetail','cx_device','cx_utmContent',
                       'cx_link','@AssetId'])
df
```

รูปที่ 3.3 ตัวอย่างชุดคำสั่งสำหรับลบคอลัมน์ที่ไม่ได้ใช้ในชุดข้อมูลออก

```
new_columns = [
    'Visited', 'acceptcookie', 'Pageview', 'refusecookie', 'Preview Detail',
    'Registration Member', 'Login Facebook', 'Click Search', 'Drop Lead',
    'Contact Staff', 'design_click_propertypage', 'click_banner', 'reserve_asset',
    'Click to Register', 'Debtstructure1', 'Payment Reserve', 'click_bamchoices_new_android',
    'click_register_bamvestor', 'design_click_create', 'design_click_start', 'design_click_viewasset',
    'paymentpaynow_qr', 'paymentpaynow_credit', 'click_bamchoices_member_android',
    'click_bamchoices_member_ios', 'click_submit_register_bamvestor', 'click_bamchoices_new_luckydraw',
    'click_bamchoices_new_applewatch', 'click_bamchoices_member_applewatch', 'click_bamchoices_new_ios'
]

def binary_value(row, col_name):
    return 1 if row['cx_event'] == col_name else 0

for col in new_columns:
    df[col] = df.apply(lambda row: binary_value(row, col), axis=1)

print(df)
```

รูปที่ 3.4 ตัวอย่างชุดคำสั่งสำหรับสร้างคอมลัมน์ใหม่จากคอลัมน์เดิมที่ต้องการ

```
df = df.drop(columns=['cx_event'])
df
```

รูปที่ 3.5 ตัวอย่างชุดคำสั่งสำหรับลบคอลัมน์ cx_event ออก

```
df = df.groupby('cx_fingerprint').agg({
    'acceptcookie': 'sum',
    'Pageview': 'sum',
    'refusecookie': 'sum',
    'Preview Detail': 'sum',
    'Registration Member': 'sum',
    'Login Facebook': 'sum',
    'Click Search': 'sum',
    'Drop Lead': 'sum',
    'Contact Staff': 'sum',
    'design_click_propertypage': 'sum',
    'click_banner': 'sum',
    'reserve_asset': 'sum',
    'Click to Register': 'sum',
    'Debtstructure1': 'sum',
    'Payment Reserve': 'sum',
    'click_bamchoices_new_android': 'sum',
    'click_register_bamvestor': 'sum',
    'design_click_create': 'sum',
    'design_click_start': 'sum',
    'design_click_viewasset': 'sum',
    'paymentpaynow_qr': 'sum',
    'paymentpaynow_credit': 'sum',
    'click_bamchoices_member_android': 'sum',
    'click_bamchoices_member_ios': 'sum',
    'click_submit_register_bamvestor': 'sum',
    'click_bamchoices_new_luckydraw': 'sum',
    'click_bamchoices_new_applewatch': 'sum',
    'click_bamchoices_member_applewatch': 'sum',
    'click_bamchoices_new_ios': 'sum'
}).reset_index()
```

รูปที่ 3.6 ตัวอย่างชุดคำสั่งสำหรับนับจำนวนข้อมูลที่เป็น Action

จากรูปที่ 3.6 จะต้องทำการรวมข้อมูลทุก Action ที่เกี่ยวข้องกับลูกค้าแต่ละคนให้อยู่ในแถวเดียว (one record per person) เพื่อให้ง่ายต่อการวิเคราะห์และการสร้างโมเดล การรวมข้อมูลนี้อ้างอิงจากคอลัมน์ cx_fingerprint ซึ่งทำให้การจับคู่ข้อมูลการกระทำ (Actions) ต่างๆ ที่เกิดจากบุคคลคนเดียวกันในแถวเดียวได้ เช่น การคลิกหน้าเว็บ การเพิ่มสินค้าในรถเข็น หรือการสั่งซื้อสินค้า โดยการรวมนี้จะใช้การสรุปหรือการรวมค่าทางสถิติ เช่น นับจำนวน Action (Count) หาค่าเฉลี่ยของ Action (Mean) หรือการเข้ารหัส Action ว่าเคยทำหรือไม่ (1/0 Encoding) วิธีนี้ช่วยลดความซับซ้อนของข้อมูลและทำให้สามารถุมุ่งเน้นที่พฤติกรรมโดยรวมของลูกค้าแทนการวิเคราะห์ในระดับ Action เดียว ซึ่งเหมาะสมสำหรับการทำโมเดลเพื่อการพยากรณ์หรือการวิเคราะห์เชิงลึก เช่น การวิเคราะห์ RFM หรือการพยากรณ์ความเป็นไปได้ในการตัดสินใจซื้อ นอกจากนี้ การรวมข้อมูลยังช่วยลดความซ้ำซ้อนและทำให้ประหยัดทรัพยากรในกระบวนการประมวลผลข้อมูล ทำให้ได้ข้อมูลที่มีความพร้อมสำหรับการนำไปวิเคราะห์ในขั้นตอนต่อไป

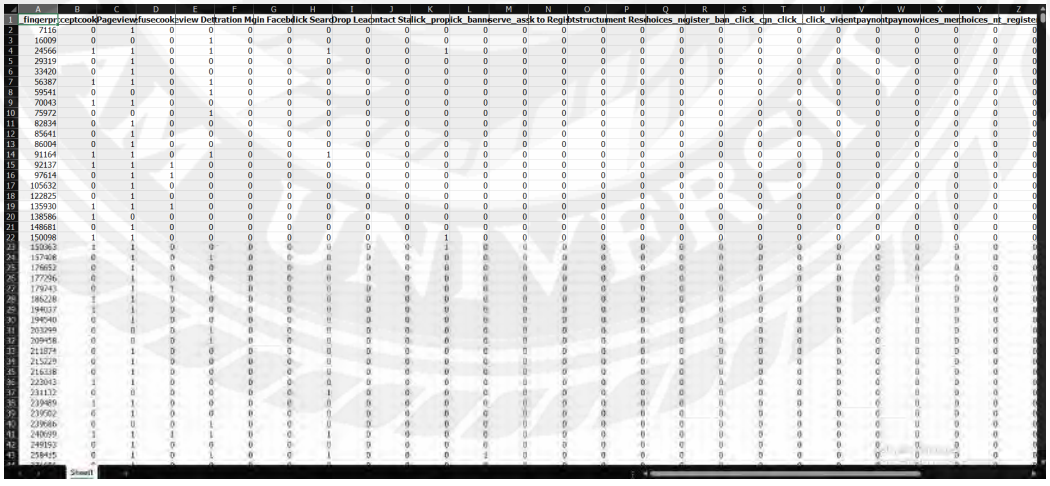
```
columns_to_transform = ['acceptcookie', 'Pageview', 'refusecookie', 'Preview Detail',
                        'Registration Member', 'Login Facebook', 'Click Search', 'Drop Lead',
                        'Contact Staff', 'design_click_propertypage', 'click_banner', 'reserve_asset',
                        'Click to Register', 'Debtstructure1', 'Payment Reserve', 'click_bamchoices_new_android',
                        'click_register_bamvestor', 'design_click_create', 'design_click_start', 'design_click_viewasset',
                        'paymentpaynow_qr', 'paymentpaynow_credit', 'click_bamchoices_member_android',
                        'click_bamchoices_member_ios', 'click_submit_register_bamvestor', 'click_bamchoices_new_luckydraw',
                        'click_bamchoices_new_applewatch', 'click_bamchoices_member_applewatch', 'click_bamchoices_new_ios']


for column in columns_to_transform:
    df[column] = df[column].apply(lambda x: 1 if x >= 1 else 0)

df
```

รูปที่ 3.7 ตัวอย่างชุดคำสั่งสำหรับการแปลงค่าจำนวนตัวเลขให้เป็น 1 และ 0

จากรูปที่ 3.7 การแปลงค่าให้เป็นบลูรีน คือ 1 กับ 0 นั้น เพื่อแสดงว่า Action นั้นมีการกระทำหรือไม่ ตัวอย่างเช่น ถ้าลูกค้ามีการกระทำ Action นั้นค่าจะเป็น 1 แต่ถ้าไม่ได้มีการกระทำเกิดขึ้นค่าจะเป็น 0 เพื่อทำให้ข้อมูลอยู่ในรูปแบบและช่วงของข้อมูลเดียวกันซึ่งจะทำให้การตัดสินใจหรือการคาดการณ์มีความแม่นยำมากขึ้น ดังนั้น การแปลงค่าให้เป็น 1 และ 0 ช่วยลดความเอนเอียงของข้อมูล (Reducing Unbalancing) ทำให้สามารถวิเคราะห์การกระทำหรือไม่กระทำพฤติกรรมของลูกค้าได้อย่างตรงประเด็น และยังทำให้โมเดลการวิเคราะห์ง่ายต่อการประมวลผลมากยิ่งขึ้น โดยเฉพาะในกรณีที่ต้องใช้โมเดลแมชชีนเลิร์นนิง ซึ่งจะทำให้ได้ผลลัพธ์ที่มีประสิทธิภาพและเชื่อถือได้มากขึ้น อีกทั้งการทำเช่นนี้ยังช่วยให้การตีความผลลัพธ์ง่ายขึ้น เช่น ถ้าโมเดลแสดงค่าความสัมพันธ์ที่สูงระหว่าง Action ที่เป็น 1 กับการซื้อสินค้า ก็สามารถระบุได้ชัดเจนว่าพฤติกรรมดังกล่าวมีผลต่อการตัดสินใจซื้อทรัพย์สินของลูกค้า



รูปที่ 3.8 ตัวอย่างข้อมูลที่ทำการ Cleansing เรียบร้อยแล้ว

3.2.4 หาความสัมพันธ์ระหว่างข้อมูล (Data Correlation)

นำข้อมูลที่ได้จัดเตรียมไว้มาทำการหาความสัมพันธ์ระหว่างข้อมูลว่า Action ใดที่มีผลต่อการตัดสินใจซื้อทรัพย์สินของลูกค้า (Drop Lead) โดยการวิเคราะห์ความสัมพันธ์ระหว่างข้อมูล (Data Correlation) โดยใช้วิธีการทางสถิติ เช่น การคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) เพื่อวัดระดับความสัมพันธ์ระหว่างตัวแปร Action และการตัดสินใจซื้อทรัพย์สินของลูกค้า (Drop Lead) ในกรณีนี้ ค่าความสัมพันธ์ (Correlation Value) ที่ได้จะมีค่าอยู่ในระหว่าง -1 และ 1 โดยที่

- ถ้าค่าความสัมพันธ์เข้าใกล้ 1 หมายถึง ตัวแปรทั้งสองมีความสัมพันธ์เชิงบวกที่แข็งแรง (Strong Positive Correlation) เช่น ยิ่งลูกค้ากระทำ Action มากเท่าใด โอกาสในการ Drop Lead ก็ยิ่งสูงขึ้น
- ถ้าค่าความสัมพันธ์เข้าใกล้ -1 หมายถึง ตัวแปรมีความสัมพันธ์เชิงลบที่แข็งแรง (Strong Negative Correlation) เช่น ยิ่งลูกค้ากระทำ Action น้อย โอกาสในการ Drop Lead ยิ่งเพิ่มขึ้น
- ถ้าค่าความสัมพันธ์เข้าใกล้ 0 หมายถึง ไม่มีความสัมพันธ์ที่ชัดเจนระหว่างตัวแปร

```
import matplotlib.pyplot as plt
import seaborn as sns

df = df.drop(columns=['cx_fingerprint'])

drop_lead_corr = df.corr()['Drop Lead'].sort_values(ascending=False)

print(drop_lead_corr)
```

รูปที่ 3.9 ตัวอย่างชุดคำสั่งสำหรับการหาค่าความสัมพันธ์ระหว่างข้อมูล

```
Drop Lead                           1.000000
Contact Staff                       0.459350
design_click_propertypage           0.163659
Login Facebook                      0.162507
reserve_asset                       0.159587
Click to Register                   0.151953
Click Search                        0.119383
design_click_create                 0.117900
Registration Member                 0.117458
design_click_start                  0.106549
click_banner                        0.106044
refusecookie                        0.093378
acceptcookie                        0.089976
Preview Detail                      0.085403
Debtstructure1                      0.076798
click_register_bamvestor            0.057804
click_bamchoices_member_ios         0.049651
paymentpaynow_qr                    0.042378
Payment Reserve                     0.037334
Pageview                            0.028893
design_click_viewasset              0.025462
click_bamchoices_new_ios            0.021127
click_bamchoices_new_luckydraw      0.016096
click_bamchoices_new_android        0.015893
paymentpaynow_credit                0.015478
click_bamchoices_member_android     0.015337
click_bamchoices_member_applewatch  0.010165
click_submit_register_bamvestor    -0.000123
click_bamchoices_new_applewatch    -0.000475
Name: Drop Lead, dtype: float64
```

รูปที่ 3.10 ตัวอย่างผลลัพท์ของการหาค่าความสัมพันธ์ระหว่างตัวแปร

จากรูปที่ 3.10 เป็นตัวอย่างผลลัพธ์ของการวิเคราะห์ความสัมพันธ์ระหว่าง Action ต่างๆ กับการตัดสินใจซื้อทรัพย์สินของลูกค้า (Drop Lead) โดยค่าความสัมพันธ์ที่แสดงมีตั้งแต่ 1.0 (ความสัมพันธ์สูงสุด) ไปจนถึงค่าที่ใกล้ 0 หรือค่าลบที่แสดงถึงความสัมพันธ์ต่ำ จะเห็นได้ว่า Action "Contact Staff" มีค่าความสัมพันธ์สูงที่สุดมีค่าเท่ากับ 0.459350 รองลงมาคือ Action "design_click_propertypage" มีค่าเท่ากับ 0.163659 และ Action "Login Facebook" มีค่าเท่ากับ 0.162507 ซึ่งแสดงว่า Action เหล่านี้มีผลต่อการตัดสินใจซื้อของลูกค้ามากกว่า Action อื่นๆ การที่ลูกค้าติดต่อเจ้าหน้าที่หรือดูหน้า Property Page บ่งบอกถึงความสนใจที่สูงขึ้น และอาจนำไปสู่การตัดสินใจซื้อได้ในทางตรงกันข้าม Action ที่มีค่าความสัมพันธ์ใกล้ 0 หรือมีค่าน้อยมาก เช่น Action "click_bamchoices_new_ios" มีค่าเท่ากับ 0.021127 หรือ Action "click_bamchoices_new_applewatch" มีค่าเท่ากับ -0.000475 แสดงว่า Action เหล่านี้มีผลต่อการตัดสินใจซื้อน้อยหรือแทบจะไม่มีผลต่อการตัดสินใจซื้อของลูกค้าเลย

ข้อสังเกตเพิ่มเติม:

- ค่า Correlation ที่เป็นบวกแสดงถึงความสัมพันธ์เชิงบวก เช่น ยิ่งมีการทำ Action มากขึ้น โอกาสในการ Drop Lead ก็ยิ่งเพิ่มสูงขึ้น
- ค่า Correlation ที่เป็นลบ เช่น -0.000475 อาจแสดงถึงความสัมพันธ์เชิงลบเล็กน้อย แต่ในกรณีนี้ ค่ามีขนาดเล็กมากจนถือได้ว่าไม่มีนัยสำคัญ

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix

X = df.drop('Drop Lead', axis=1)
y = df['Drop Lead']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

print("Accuracy score:", model.score(X_test, y_test))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))
```

รูปที่ 3.11 ตัวอย่างการสร้างโมเดลเพื่อการทำนายด้วย Random Forest

จากรูปที่ 3.11 แสดงขั้นตอนการสร้างโมเดลเพื่อการทำนายด้วยโมเดล Random Forest ซึ่งเป็นหนึ่งในอัลกอริทึมที่นิยมใช้สำหรับการจัดประเภทของกลุ่มข้อมูล (Classification) โดยใช้แนวคิดการรวมต้นไม้ตัดสินใจหลายต้นเข้าด้วยกันเพื่อปรับปรุงความแม่นยำของการทำนาย

1. การเตรียมข้อมูล:
   - ลบคอลัมน์ Drop Lead ออกจาก X ซึ่งเป็นฟีเจอร์อินพุต (features) และเก็บ Drop Lead ไว้ใน y ซึ่งเป็นข้อมูลเป้าหมาย (target).
   - แบ่งชุดข้อมูลออกเป็นชุดข้อมูลสำหรับการฝึก (Training Dataset) และชุดข้อมูลสำหรับการทดสอบ (Testing Dataset) โดยกำหนดขนาดชุดทดสอบเป็น 20% (test_size=0.2) และตั้งค่าการสุ่มด้วย random_state=42
2. การสร้างและฝึกโมเดล:
   - ใช้ Random Forest โดยกำหนดค่า random_state=42 เพื่อสร้างโมเดล
   - ฝึกสอนโมเดลด้วยชุดข้อมูลฝึกหรือชุดข้อมูลผู้สอน (X_train และ y_train).
3. การทำนายผลลัพธ์:
   - ใช้โมเดลที่ฝึกแล้วทำการทำนาย (Predict) ชุดข้อมูลทดสอบ (X_test).
4. การประเมินผลลัพธ์:
   - คำนวณคะแนนความแม่นยำ (Accuracy score) ด้วย model.score.
   - แสดงรายงานการจัดประเภท (Classification Report) และเมทริกซ์ความสับสน (Confusion Matrix) เพื่อตรวจสอบผลลัพธ์ในเชิงลึก

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.preprocessing import LabelEncoder
import matplotlib.pyplot as plt
from sklearn import tree

label_encoder = LabelEncoder()
df['Drop Lead'] = label_encoder.fit_transform(df['Drop Lead'])

X = df.drop('Drop Lead', axis=1)
y = df['Drop Lead']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = DecisionTreeClassifier(random_state=42)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

print("Accuracy score:", model.score(X_test, y_test))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))

plt.figure(figsize=(100,40))
tree.plot_tree(model, filled=True, feature_names=X.columns, class_names=['No', 'Yes'], rounded=True, fontsize=12)
plt.title("Decision Tree for Drop Lead")
plt.show()
```

รูปที่ 3.12 ตัวอย่างการสร้างโมเดลการทำนายด้วย Decision Tree

จากรูปที่ 3.12 แสดงการสร้างโมเดลด้วย Decision Tree ซึ่งใช้โครงสร้างต้นไม้สำหรับการตัดสินใจ โดยมีขั้นตอนที่คล้ายคลึงกับการสร้างโมเดล Random Forest แต่มีการเพิ่มการแปลงข้อมูลและการแสดงผล

1. การเตรียมข้อมูล:
   - ใช้ LabelEncoder เพื่อแปลงข้อมูลใน Drop Lead ให้อยู่ในรูปแบบตัวเลขก่อนนำไปสร้างโมเดล.
   - ลบคอลัมน์ Drop Lead จาก X และเก็บข้อมูลเป้าหมายใน y.
   - แบ่งชุดข้อมูลเป็นชุดฝึกและชุดทดสอบเช่นเดียวกับรูปที่ 3.11.
2. การสร้างและฝึกโมเดล:
   - ใช้ DecisionTreeClassifier โดยตั้งค่า random_state=42 เพื่อสร้างโมเดล.
   - ฝึกโมเดลด้วย X_train และ y_train.
3. การทำนายผลลัพธ์:
   - ทำการพยากรณ์ด้วยชุดข้อมูลทดสอบ (X_test).
4. การประเมินผลลัพธ์:
   - คำนวณคะแนนความแม่นยำ (Accuracy score) และแสดงผล Classification Report และ Confusion Matrix.
5. การแสดงโครงสร้างต้นไม้:
   - ใช้ tree.plot_tree เพื่อวาดโครงสร้างของต้นไม้ตัดสินใจ โดยระบุฟีเจอร์ที่ใช้ (feature_names) และค่าของคลาส (class_names เช่น Yes และ No).

- กำหนดลักษณะของกราฟ เช่น สีสัน (filled=True), ขนาดตัวอักษร (fontsize=12), และขนาดภาพ (figsize=(100, 40)).

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import classification_report, confusion_matrix

# กำหนดค่า random_state เป็น None เพื่อสุ่มใหม่ทุกครั้ง
random_state = None

# แบ่งข้อมูล X และ y
X = df.drop('Drop Lead', axis=1)
y = df['Drop Lead']

# แบ่งข้อมูล train และ test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# 1. Logistic Regression

logistic_model = LogisticRegression(random_state=42)
logistic_model.fit(X_train, y_train)
y_pred_logistic = logistic_model.predict(X_test)
print("Logistic Regression:")
print("Accuracy score:", logistic_model.score(X_test, y_test))
print("\nClassification Report:\n", classification_report(y_test, y_pred_logistic))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred_logistic))

# 2. Support Vector Machine (SVM)

svm_model = SVC(random_state=42)
svm_model.fit(X_train, y_train)
y_pred_svm = svm_model.predict(X_test)
print("\nSupport Vector Machine (SVM):")
print("Accuracy score:", svm_model.score(X_test, y_test))
print("\nClassification Report:\n", classification_report(y_test, y_pred_svm))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred_svm))
```

รูปที่ 3.13 ตัวอย่างการสร้างโมเดลการทำนายด้วย Logistic Regression และ Support Vector Machine (SVM)

จากรูปที่ 3.13 แสดงการสร้างโมเดลเพื่อการทำนายด้วยโมเดล Logistic Regression และ SVM โดยใช้ข้อมูลสำหรับการจัดประเภท (Classification) และมีขั้นตอนดังนี้

1. การเตรียมข้อมูล
   - ข้อมูลเป้าหมาย (Drop Lead) ถูกแยกออกจากฟีเจอร์ (features) อื่น ๆ และหากข้อมูลเป้าหมายอยู่ในรูปแบบข้อความ เช่น "Yes" และ "No" จะถูกแปลงเป็นตัวเลขเพื่อให้สามารถใช้งานกับโมเดลได้.
   - ข้อมูลถูกแบ่งออกเป็นชุดข้อมูลสำหรับการฝึก (training set) และการทดสอบ (test set) โดยชุดข้อมูลทดสอบมีสัดส่วน 20% ของข้อมูลทั้งหมด

2. Logistic Regression
   - Logistic Regression เป็นโมเดลที่เหมาะสำหรับการจัดประเภทข้อมูลที่สามารถแยกกันด้วยเส้นตรง (linearly separable). โมเดลนี้ใช้อัลกอริทึมการถดถอยเชิงลอจิสติก (logistic regression) และฟังก์ชัน sigmoid เพื่อคำนวณความน่าจะเป็นของแต่ละคลาส.
   - หลังจากการฝึกโมเดลแล้ว โมเดลจะถูกใช้ในการพยากรณ์ข้อมูลในชุดทดสอบ และผลลัพธ์จะถูกวัดโดยค่าความแม่นยำ (Accuracy), รายงานการจัดประเภท (Classification Report), และเมทริกซ์ความสับสน (Confusion Matrix).
3. Support Vector Machine
   - SVM เป็นโมเดลที่ใช้เทคนิค Support Vector Machine (SVM) ซึ่งมุ่งเน้นการสร้างไฮเปอร์เพลน (hyperplane) ที่เหมาะสมที่สุดในการแยกกลุ่มข้อมูล.
   - หากข้อมูลไม่สามารถแยกกันด้วยเส้นตรงได้ โมเดล SVC สามารถใช้ kernel trick (เช่น linear, rbf) เพื่อจับรูปแบบที่ซับซ้อนในข้อมูล.
   - โมเดลนี้ถูกประเมินผลในลักษณะเดียวกับ Logistic Regression โดยใช้ค่าความแม่นยำ รายงานการจัดประเภท และเมทริกซ์ความสับสน.

บทที่ 4
การนำเสนอแผนภาพของข้อมูล

หลังจากที่ได้เตรียมข้อมูล (Data Preparation) จนได้ข้อมูลที่สามารถนำไปทำการวิเคราะห์เพื่อศึกษาพฤติกรรมใดที่ส่งผลต่อการตัดสินใจซื้อของลูกค้า (Data Correlation) และการหาโมเดลการทำนายผล (Data Prediction) ตามเงื่อนไขที่กำหนดได้แล้ว ต่อไปจะเป็นการนำผลลัพธ์ที่ได้มานำเสนอให้ผู้ใช้เข้าใจได้ง่าย และสามารถเข้าถึงข้อมูลเชิงลึกได้ง่าย นั่นคือการนำเสนอแผนภาพของข้อมูล (Data Visualization) โดยผู้จัดทำได้เลือกการนำเสนอผลในรูปแบบของกราฟ (Graph) และแผนภูมิแท่ง (Bar Chart) และกราฟแบบอื่นๆ ตามความเหมาะสมกับผลลัพธ์ที่ได้

4.1 การวิเคราะห์พฤติกรรมที่มีผลต่อการตัดสินใจซื้อทรัพย์สินของลูกค้า (Drop Lead)

ในการนำเสนอผลลัพธ์จากการวิเคราะห์ความสัมพันธ์ระหว่างข้อมูล (Correlation) เพื่อหาพฤติกรรมที่ส่งผลต่อการตัดสินใจซื้อทรัพย์สินของลูกค้า (Drop Lead) สามารถใช้กราฟหรือแผนภูมิที่เหมาะสมเพื่อช่วยให้เข้าใจข้อมูลได้ง่ายขึ้น โดยในกรณีนี้ ค่าความสัมพันธ์จากภาพที่แสดงเป็นตัวเลข ได้ผลลัพธ์ดังนี้

```
Drop Lead                           1.000000
Contact Staff                       0.459350
design_click_propertypage           0.163659
Login Facebook                      0.162507
reserve_asset                       0.159587
Click to Register                   0.151953
Click Search                        0.119383
design_click_create                 0.117900
Registration Member                 0.117458
design_click_start                  0.106549
click_banner                        0.106044
refusecookie                        0.093378
acceptcookie                        0.089976
Preview Detail                      0.085403
Debtstructure1                      0.076798
click_register_bamvestor            0.057804
click_bamchoices_member_ios         0.049651
paymentpaynow_qr                    0.042378
Payment Reserve                     0.037334
Pageview                            0.028893
design_click_viewasset              0.025462
click_bamchoices_new_ios            0.021127
click_bamchoices_new_luckydraw      0.016096
click_bamchoices_new_android        0.015893
paymentpaynow_credit                0.015478
click_bamchoices_member_android     0.015337
click_bamchoices_member_applewatch  0.010165
click_submit_register_bamvestor    -0.000123
click_bamchoices_new_applewatch    -0.000475
Name: Drop Lead, dtype: float64
```

รูปที่ 4.1 ผลลัพธ์ของการทำ Data Correlation

จากรูปที่ 4.1 การวิเคราะห์ความสัมพันธ์ พบว่าพฤติกรรมที่มีความสัมพันธ์กับการ Drop Lead (การตัดสินใจซื้อ) มากที่สุดคือ

- Contact Staff มีค่าสูงสุดเท่ากับ 0.459 แสดงให้เห็นว่าการติดต่อเจ้าหน้าที่ มีอิทธิพลอย่างมากต่อการตัดสินใจซื้อทรัพย์สินของลูกค้ามากที่สุด
- design_click_propertypage มีค่าเท่ากับ 0.163 การคลิกลิงก์ที่เกี่ยวกับหน้าอสังหาริมทรัพย์มีอิทธิพลต่อการตัดสินใจซื้อเป็นอันดับที่ 2
- Login Facebook มีค่าเท่ากับ 0.162 การเข้าสู่ระบบด้วย Facebook มีผลต่อการตัดสินใจซื้อเป็นลำดับที่ 3
- reserve_asset มีค่าเท่ากับ 0.159 และ Click to Register มีค่าเท่ากับ 0.151 พฤติกรรมการสำรองทรัพย์สินและการลงทะเบียนก็มีอิทธิพลเป็นลำดับที่ 4 ทั้ง 2 การกระทำ

พฤติกรรมอื่นๆ ที่มีความสัมพันธ์ระดับปานกลางถึงต่ำ สามารถใช้ประกอบการวิเคราะห์เพิ่มเติมเพื่อหาแนวโน้มที่ละเอียดขึ้นได้ดังนี้

- click_submit_register_bamvestor และ click_bamchoices_new_new_applewatch มีค่าสัมประสิทธิ์เป็นลบ ซึ่งบ่งบอกถึงความสัมพันธ์ในทางตรงกันข้ามต่อการตัดสินใจซื้อ

4.2 การสร้างโมเดลเพื่อการทำนายด้วย Random Forest

ในการนำเสนอผลลัพธ์จากโมเดล Random Forest จะแสดงให้เห็นผลลัพธ์การทำนายที่โมเดลตัวนี้ทำได้ เพื่อประกอบการตัดสินใจว่าข้อมูลตัวนี้เหมาะกับโมเดลชนิดไหนที่ให้ค่าความแม่นยำสูงสุด

```
Accuracy score: 0.9936521781266657

Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00     82030
           1       0.47      0.09      0.15       518

    accuracy                           0.99     82548
   macro avg       0.73      0.55      0.58     82548
weighted avg       0.99      0.99      0.99     82548


Confusion Matrix:
 [[81976    54]
 [  470    48]]
```

รูปที่ 4.2 แสดงผลลัพธ์ของโมเดล Random Forest

จากรูปที่ 4.2 แสดงผลลัพธ์ของโมเดล Random Forest โมเดลมีความแม่นยำสูงถึง 99.36% ซึ่งหมายความว่าโมเดลสามารถทำนายได้ถูกต้องในกรณีส่วนใหญ่ แต่ควรพิจารณาความแม่นยำของแต่ละคลาสเพิ่มเติม เนื่องความไม่สมดุลของชุดข้อมูลที่มีจำนวนตัวอย่างในคลาส 0 มากกว่าคลาส 1 Precision สำหรับคลาส 0 คือ 0.99 หมายถึงจากการทำนายทั้งหมดเป็นคลาส 0 มีเพียง 1%

ที่ผิดพลาด Recall สำหรับคลาส 0 คือ 1.00 แสดงว่าโมเดลสามารถจับตัวอย่างทั้งหมดของคลาส 0 ได้ครบ Precision สำหรับคลาส 1 คือ 0.47 ซึ่งต่ำ แปลว่ามีการทำนายคลาส 1 ผิดพอสมควร Recall สำหรับคลาส 1 คือ 0.09 หมายความว่าโมเดลจับตัวอย่างคลาส 1 ได้เพียง 9% เท่านั้น ซึ่งเป็นปัญหาที่พบได้เมื่อข้อมูลไม่สมดุล

4.3 การสร้างโมเดลเพื่อการทำนายด้วย Decision Tree

ในการนำเสนอผลลัพธ์จากใช้โมเดล Decision Tree จะแสดงให้เห็นผลลัพธ์การทำนายที่โมเดลตัวนี้ทำได้ เพื่อประกอบการตัดสินใจว่าข้อมูลตัวนี้เหมาะกับโมเดลชนิดไหนที่ให้ค่าความแม่นยำสูงสุด

```
Accuracy score: 0.993543150651742

Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00     82030
           1       0.43      0.08      0.14       518

    accuracy                           0.99     82548
   macro avg       0.71      0.54      0.57     82548
weighted avg       0.99      0.99      0.99     82548


Confusion Matrix:
 [[81972    58]
 [  475    43]]
```

รูปที่4.3 แสดงผลลัพธ์ของโมเดล Decision Tree

จากรูปที่ 4.3 แสดงผลลัพธ์ของโมเดล Decision Tree โมเดลมีความแม่นยำสูงถึง 99.35% ซึ่งดูเหมือนแม่นยำมาก แต่ควรพิจารณาในรายละเอียดของแต่ละคลาส เนื่องจากข้อมูลไม่สมดุล (คลาส 0 มีจำนวนมากกว่าคลาส 1 อย่างมาก) คลาส 0: Precision = 0.99 หมายถึงการทำนายคลาส 0 ถูกต้องสูงมาก Recall = 1.00 โมเดลสามารถจับตัวอย่างคลาส 0 ได้ครบทุกตัวอย่าง F1-score = 1.00 เป็นคะแนนที่สะท้อนความสมดุลระหว่าง precision และ recall คลาส 1: Precision = 0.43 หมายถึงเมื่อโมเดลทำนายเป็นคลาส 1 มีเพียง 43% ที่ถูกต้อง Recall = 0.08 โมเดลสามารถจับตัวอย่างคลาส 1 ได้เพียง 8% เท่านั้น F1-score = 0.14 บ่งบอกถึงการทำนายคลาส 1 ยังไม่ดีนัก

4.4 การสร้างโมเดลเพื่อการทำนายด้วย Logistic Regression และ Support Vector Machine

ในการนำเสนอผลลัพธ์จากใช้โมเดล Logistic Regression และ Support Vector Machine จะแสดงให้เห็นผลลัพธ์การทำนายที่โมเดลตัวนี้ทำได้ เพื่อประกอบการตัดสินใจว่าข้อมูลตัวนี้เหมาะกับโมเดลชนิดไหนที่ให้ค่าความแม่นยำสูงสุด

```
Logistic Regression:
Accuracy score: 0.9936521781266657

Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00     82030
           1       0.46      0.07      0.12       518

    accuracy                           0.99     82548
   macro avg       0.73      0.53      0.56     82548
weighted avg       0.99      0.99      0.99     82548


Confusion Matrix:
 [[81989    41]
 [  483    35]]

Support Vector Machine (SVM):
Accuracy score: 0.9936885206183069

Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00     82030
           1       0.47      0.05      0.09       518

    accuracy                           0.99     82548
   macro avg       0.73      0.52      0.54     82548
weighted avg       0.99      0.99      0.99     82548


Confusion Matrix:
 [[82002    28]
 [  493    25]]
```

รูปที่4.4 แสดงผลลัพธ์ของโมเดล Logistic Regression และ Support Vector Machine

จากรูปที่ 4.4 แสดงผลลัพธ์ของโมเดล Logistic Regression และ Support Vector Machine โมเดลมีความแม่นยำสูงถึง 99.36% ซึ่งดูเหมือนแม่นยำมาก แต่ควรพิจารณาในส่วนของโมเดล Logistic Regression รายละเอียดของแต่ละคลาส เนื่องจากข้อมูลไม่สมดุล (คลาส 0 มีจำนวนมากกว่าคลาส 1 อย่างมาก) คลาส 0: Precision = 0.99 หมายถึงการทำนายคลาส 0 ถูกต้องสูงมาก Recall = 1.00 โมเดลสามารถจับตัวอย่างคลาส 0 ได้ครบทุกตัวอย่าง F1-score = 1.00 เป็นคะแนนที่สะท้อนความสมดุลระหว่าง precision และ recall คลาส 1: Precision = 0.46 หมายถึงเมื่อโมเดลทำนายเป็นคลาส 1 มีเพียง 46% ที่ถูกต้อง Recall = 0.07 โมเดลสามารถจับตัวอย่างคลาส 1 ได้เพียง 7% เท่านั้น F1-score = 0.12 บ่งบอกถึงการทำนายคลาส 1 ยังไม่ดีนัก และในส่วนของ โมเดล Support Vector Machine รายละเอียดของแต่ละคลาส เนื่องจากข้อมูลไม่สมดุล (คลาส 0 มีจำนวนมากกว่าคลาส 1 อย่างมาก) คลาส 0: Precision = 0.99 หมายถึงการทำนายคลาส 0 ถูกต้องสูงมาก Recall = 1.00 โมเดลสามารถจับตัวอย่างคลาส 0 ได้ครบทุกตัวอย่าง F1-score = 1.00 เป็นคะแนนที่

สะท้อนความสมดุลระหว่าง precision และ recall คลาส 1: Precision = 0.47 หมายถึงเมื่อโมเดลทำนายเป็นคลาส 1 มีเพียง 47% ที่ถูกต้อง Recall = 0.05 โมเดลสามารถจับตัวอย่างคลาส 1 ได้เพียง 5% เท่านั้น F1-score = 0.9 บ่งบอกถึงการทำนายคลาส 1 ยังไม่ดีนัก

บทที่ 5
สรุปผลและข้อเสนอแนะ

5.1 สรุปผลปริญญานิพนธ์

ผลที่ได้รับจากการวิเคราะห์พฤติกรรมการเข้าใช้งานเว็บไซต์ของลูกค้าของบริษัทสินทรพัย์แห่งเหนึ่งที่มีผลต่อการตัดสินใจซื้อทรัพย์สินและการสร้างแบบจำลองเพื่อการทำนาย โดยใช้ของมูลของเดือนมกราคม - กรกฎาคม 2564 โดยนำเสนอด้วยแผนภาพข้อมูล (Data Visualization) เพื่อให้ง่ายต่อการแสดงผลลัพธ์ ซึ่งได้ผลลัพธ์จากการวิเคราะห์ดังนี้

5.1.1 สำหรับพฤติกรรมที่ส่งผลต่อการตัดสิใจซื้อ หรือ ตัวแปรที่มีค่า Correlation สูงและเป็นบวก แสดงถึงความสัมพันธ์โดยตรงกับการ Drop Lead โดยถ้าค่าตัวแปรนั้นเพิ่มขึ้น โอกาส Drop Lead ก็เพิ่มขึ้น ประกอบด้วย 1. Contact Staff (0.459350) การติดต่อเจ้าหน้าที่ มีผลอย่างมาก 2. Design Click Propertypage (0.163659) การคลิกที่หน้า Property 3. Login Facebook (0.162507) การเข้าสู่ระบบผ่าน Facebook 4. Reserve Asset (0.159587) การจองสินทรัพย์ และ 5. Click to Register (0.151953) การคลิกเพื่อสมัครสมาชิก

5.1.2 สำหรับโมเดลเพื่อการทำนายที่เลือกใช้คือ โมเดล Random Forest โดยเมื่อเปรียบเทียบค่าทั้งหมดของทั้ง 4 โมเดลที่เลือกใช้ ทุกโมเดลมีค่า Accuracy ใกล้เคียงกันมาก (>99%) ซึ่งบ่งบอกถึงความสามารถในการจัดกลุ่มข้อมูลโดยรวมได้ดี อย่างไรก็ตาม Accuracy อาจไม่ใช่ตัวชี้วัดที่เหมาะสมที่สุดสำหรับข้อมูลที่ไม่สมดุล (Imbalanced Data) เช่นในกรณีนี้ เนื่องจาก Class "1" มีจำนวนตัวอย่างน้อยกว่ามาก ส่วน Recall ของ Class "1" มีค่าต่ำมากในทุกโมเดล โดย Random Forest ทำได้ดีกว่าเล็กน้อย Precision ของ Class "1" สำหรับ Random Forest เท่ากับ Support Vector Machine ซึ่งทั้งสองโมเดลมีค่าเท่ากัน ความแม่นยำในการระบุ Class "1" มากกว่าโมเดลตัวอื่น F1-Score ของ Random Forest มีค่ามากกว่าโมเดลอื่นแต่ยังคงต่ำ ซึ่งสะท้อนปัญหาจาก Recall ต่ำ จากเหตุผลข้างต้นแสดงว่า โมเดล Random Forest มีผลลัพธ์ดีที่สุดสำหรับ 4 โมเดลที่นำมาทดสอบ และ Random Forest มีความสามารถในการลด Overfitting ได้ดีกว่า Decision Tree เนื่องจากใช้การผสมของหลายต้นไม้ (Ensemble Learning) ความเสถียรสูงกว่า Random Forest ใช้การผลเฉลี่ยจากต้นไม้หลายต้นใน Decision Tree จึงลดผลกระทบจากข้อมูลที่ผิดปกติ (Noise)

5.2 ข้อเสนอแนะ

5.2.1 บริษัทสินทรัพย์อาจนำผลลัพธ์จากการวิเคราะห์พฤติกรรมที่มีผลต่อการตัดสินใจซื้อทรัพย์สิน (Action) ไปใช้ในการหากลุ่มเป้าหมาย บริษัทควรมุ่งเน้นการปรับปรุงหรือเพิ่มความสำคัญของ Action ที่มีผลต่อการตัดสินใจซื้อทรัพย์สิน เช่น การปรับปรุงระบบ Contact Staff ให้มีประสิทธิภาพมากยิ่งขึ้น หรือเพิ่มความน่าสนใจของฟีเจอร์ Reserve Asset เพื่อเพิ่มโอกาสปิดการขาย หรือใช้ผลวิเคราะห์ในการออกแบบแคมเปญการตลาดเฉพาะบุคคล (Personalized Marketing) เพื่อดึงดูดลูกค้ากลุ่มที่แสดงพฤติกรรมใกล้เคียงกับ Action ที่มีผลสูงสุด

5.2.2 เพื่อให้ผลวิเคราะห์แม่นยำมากขึ้น ควรปรับปรุงความสมดุลของข้อมูล (Data Balancing) ด้วยเทคนิค เช่น Oversampling หรือ Undersampling สำหรับ Class ที่มีจำนวนน้อย รวบรวมข้อมูลเพิ่มเติม เช่น ข้อมูลเชิงพฤติกรรมในช่องทางอื่นๆ (Customer Journey Data) เพื่อนำมาเสริมสร้างมุมมองที่รอบด้านยิ่งขึ้น

5.2.3 การปรับปรุงประสิทธิภาพและความสามารถในการใช้งาน บริษัทอาจพัฒนาระบบแนะนำสินค้าหรือบริการที่เหมาะสมสำหรับลูกค้าแต่ละราย โดยอิงจากผลวิเคราะห์พฤติกรรมที่มีผลต่อการตัดสินใจซื้อทรัพย์สิน ติดตามผลการเปลี่ยนแปลงในพฤติกรรมของลูกค้าอย่างต่อเนื่อง เพื่อให้การวิเคราะห์สามารถตอบสนองต่อความต้องการที่เปลี่ยนแปลงไป จัดอบรมทีมงานเกี่ยวกับข้อมูลและพฤติกรรมลูกค้า เพื่อเพิ่มประสิทธิภาพในการสื่อสารและให้บริการแก่ลูกค้า

บรรณานุกรม

แคโรไลน์. (2567). *Data Understanding ขั้นตอนที่หลายคนมองข้าม.* https://www.coraline.co.th/single-post/data-understanding-process

ชัยภพ แจ่มจำรัส. (2565, 15 กันยายน). *รู้จักกับ Decision Tree มันคือต้นไม้อะไร ทำงานอย่างไร?.* [เว็บบล็อก]. https://www.borntodev.com/2022/09/15/รู้จักกับ-decision-tree/

ดิทโต้. (2566). *ทำความรู้จักการเก็บรวบรวมข้อมูล หรือ Data Collection คือ อะไร.* https://www.dittothailand.com/dittonews/gov-what-is-data-collection/

นรุตน์ สุนทรานนท์. (2566, 15 ตุลาคม). *SVM คือ อะไร.* [เว็บบล็อก]. https://www.nerd-data.com/svm/

พนาเอก. (2557, 11 มิถุนายน). *การทำ data preparation อย่างมืออาชีพ.* [เว็บบล็อก]. https://bzinsight.wordpress.com/2014/06/11/การทำ-data-preparation-อย่างมืออาชีพ/

มีเดียม. (2562, 29 กันยายน). *การวิเคราะห์ความสัมพันธ์กับข้อมูลขนาดใหญ่.* [เว็บบล็อก]. https://medium.com/@pradyasin/random-forest-คืออะไร-74d2a0af3d7

วรพิชญา ระเบียบโลก. (2564, 8 กรกฎาคม). *การวิเคราะห์ความสัมพันธ์กับข้อมูลขนาดใหญ่* . [เว็บบล็อก]. https://bdi.or.th/big-data-101/correlation-analysis-in-big-data/

วิกิพีเดีย. (ม.ป.ป). *วิทยาการข้อมูล.* วันที่สืบค้น 3 พฤศจิกายน 2567, จาก https://th.wikipedia.org/wiki/วิทยาการข้อมูล

เอดับเบิ้ลยูเอส. (2567). *รีเกรสชันโลจิสติกคืออะไร.* https://aws.amazon.com/th/what-is/logistic-regression/

เอดับเบิ้ลยูเอส. (2567). *Python คืออะไร.* https://aws.amazon.com/th/what-is/python/

แอฟฟินิตี้. (2563). *Data Analytics คือ อะไร?.* https://affinity.co.th/data-analytics/