การจัดกลุ่มลูกค้าบัตรเครดิตด้วยเทคนิคเค-มีนส์และต้นไม้ตัดสินใจเพื่อแนะนำแผนการตลาดเฉพาะกลุ่ม

ฐิติพร ฐิติพรธรรม[1*], ศุภร คนธภักดี[2]

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการแบ่งกลุ่มลูกค้าของบริษัทบัตรเครดิตแห่งหนึ่ง โดยการศึกษาแบ่งออกเป็น 2 ช่วง คือ ช่วงที่หนึ่งการแบ่งกลุ่มด้วยแบบจำลองเค-มีนส์ และช่วงที่สองการทำนายกลุ่มด้วยแบบจำลองต้นไม้ตัดสินใจ โดยในช่วงแรกใช้วิธีการหาจำนวนกลุ่ม (คลัสเตอร์) ที่เหมาะสมด้วย วิธีการ Elbow method, Davies Bouldin index และCalinski Harabasz index และในช่วงที่สองให้แบบจำลองต้นไม้ตัดสินใจทำนายกลุ่มลูกค้า เพื่อหาฟีเจอร์สำคัญที่ส่งผลให้แบบจำลองใช้เป็นกฎการตัดสินใจในการแบ่งกลุ่มลูกค้า และแสดงถึงลักษณะของลูกค้าในแต่ละกลุ่ม โดยการศึกษาในครั้งนี้ช่วยให้เข้าใจปัจจัยที่มีอิทธิพลต่อพฤติกรรมและความชอบของลูกค้าในแต่ละกลุ่มได้อย่างชัดเจนยิ่งขึ้น และในท้ายที่สุดงานวิจัยนี้มีการนำเสนอแนวทางสำหรับการออกแคมเปญทางการตลาด และโปรโมชั่นเพื่อตอบสนองความต้องการของลูกค้าในแต่ละกลุ่ม

คำสำคัญ : แบบจำลองเค-มีนส์, แบบจำลองต้นไม้ตัดสินใจ, การจัดกลุ่มลูกค้า,การวิเคราะห์ซิลูเอต, ตัวชี้วัดเดวี-บุลดีน, ตัวชี้คาลินสกี-ฮาราบาซ

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 085-9656632 E-mail address: thitiporn.mmo@g.swu.ac.th

K-Means Clustering and Decision Tree Classification Techniques for Credit Card Customer Segmentation and Personalized Marketing

Thitiporn Thitiporndharma[1*], Subhorn Khontapagdee[2]

Abstract

This study aims to analyze customer segmentation of a credit card company. The analysis consists of two phases: K-Means clustering and Decision Tree classification. In the initial phase, various numbers of clusters were explored using Elbow method, Silhouette analysis, Davies-Bouldin index, and Calinski-Harabasz index. The second phase focused on Decision Tree classification to identify key features that differentiate customer groups and capture characteristics of each cluster. Finally, guidelines for developing customized campaigns ore promotions were provided.

Keywords : K-means clustering, Decision Tree classification, Customer segmentation, Silhouette analysis, Davies-Bouldin index, Calinski-Harabasz index.

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author:Tel.: 085-9656632 E-mail address: thitiporn.mmo@g.swu.ac.th

บทนำ

จากการเติบโตของเทคโนโลยีในปัจจุบัน การแข่งขันทางธุรกิจที่สูงขึ้น รวมถึงเป็นยุคที่ลูกค้าสามารถเข้าถึงข้อมูลของสินค้า และบริการของแบรนด์ต่างๆได้อย่างง่ายดายด้วยโทรศัพท์มือถือเพียงเครื่องเดียว ทำให้แต่ละธุรกิจต่างต้องแข่งขันกันอย่างหนัก เพื่อช่วงชิงลูกค้า และส่วนแบ่งการตลาด จากสาเหตุเหล่านี้ ทำให้แต่ละธุรกิจมีการวางแผนกลยุทธการตลาด ทั้งในด้านสินค้า บริการ ราคา และแคมเปญทางการตลาดในรูปแบบต่างๆ เพื่อนำเสนอสินค้า หรือ บริการ ให้ตรงกับความต้องการของลูกค้า ทั้งนี้การส่งเสริมการขายที่ธุรกิจนิยมใช้กันบ่อยๆ คือ การลดแลกแจกแถม ซึ่งเป็นแคมเปญการตลาดที่ส่ง ผลดีในระยะสั้น แต่ในระยะยาวไม่ส่งผลดีมากนัก เนื่องจากธุรกิจจะไม่สามารถรักษาลูกค้าที่ซื้อสินค้าเฉพาะช่วงลดราคาไว้ได้ และลูกค้าสามารถเปลี่ยนใจไปซื้อสินค้าจากคู่แข่งได้เสมอ หากสินค้าของคู่แข่งมีราคาที่ต่ำกว่า ดังนั้น ในการวางแผนกลยุทธทางการตลาด นอกจากการวางแผนกลยุทธในการหาลูกค้าใหม่แล้ว การรักษาลูกค้าให้อยู่กับธุรกิจก็มีความสำคัญเช่นกัน ซึ่งการที่ธุรกิจจะสามารถรักษาลูกค้าให้อยู่กับธุรกิจได้นั้น สิ่งที่สำคัญ คือ การรู้ข้อมูลเชิงลึกของลูกค้า เพื่อให้ธุรกิจสามารถนำเสนอสินค้า บริการ หรือ ข้อเสนอพิเศษที่ตรงกับความต้องการของลูกค้าของธุรกิจเอง อย่างถูกที่ ถูกเวลา และสร้างความแตกต่างจากคู่แข่ง

อนึ่ง การที่ธุรกิจรู้จักลูกค้า และสื่อสารหาลูกค้าทุกคนด้วยข้อเสนอแบบเดียวกัน อาจไม่ใช่วิธีการที่เหมาะสม ก่อให้เกิดต้นทุนสูง และไม่ได้ผลตอบรับที่ดี เนื่องจากลูกค้าแต่ละรายมีความต้องการที่แตกต่างกันออกไป ในขณะเดียวกันการสื่อสารหาลูกค้าด้วยรูปแบบ และข้อเสนอที่เฉพาะเจาะจงลงไปในแต่ละบุคคล ในทางปฏิบัติค่อนข้างทำได้ยาก โดยเฉพาะธุรกิจที่มีขนาดกลางถึงขนาดใหญ่ซึ่งมีปริมาณลูกค้าจำนวนมาก เพราะต้องใช้ต้นทุนทางด้านงบประมาณ แรงงาน รวมถึงต้นทุนเวลา ในการสื่อสารหาลูกค้าให้ครบทุกราย ในทางการตลาด ใช้เทคนิคการแบ่งลูกค้าออกเป็นกลุ่มย่อย โดยจัดกลุ่มลูกค้าที่มีลักษณะคล้ายคลึงกันอยู่ในกลุ่มเดียวกัน เพื่อให้ธุรกิจสามารถเข้าใจลักษณะเฉพาะที่มีร่วมกันของลูกค้า เข้าใจพฤติกรรมในการใช้จ่าย

การจัดกลุ่มลูกค้าเป็นหัวใจสำคัญของกลยุทธทางการตลาดซึ่งเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (Unsupervised Machine Learning) โดยเฉพาะอัลกอริทึมการจัดกลุ่ม (Clustering Algorithm) มีประสิทธิภาพในการจัดกลุ่มลูกค้าที่มีพฤติกรรม ความชอบ และลักษณะเฉพาะที่คล้ายคลึงกัน โดยไม่จำเป็นต้องใช้ชุดข้อมูลที่มีเลเบล

การวิเคราะห์การจัดกลุ่ม (Clustering Analysis) ในเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน ช่วยให้ธุรกิจสามารถระบุกลุ่มลูกค้าที่มีพฤติกรรมการใช้จ่าย ลักษณะทางประชากรศาสตร์ และความชอบที่คล้ายคลึงกัน การเข้าใจกลุ่มลูกค้าเหล่านี้ทำให้ธุรกิจสามารถพัฒนาแคมเปญทางการตลาดแบบเฉพาะเจาะจงที่ตรงกับความต้องการของลูกค้าในแต่ละกลุ่ม

แม้ว่าเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอนอย่างอัลกอริทึมการจัดกลุ่มจะมีประสิทธิภาพในการจัดกลุ่มลูกค้าที่มีพฤติกรรมและความชอบที่คล้ายคลึงกัน แต่เทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอนก็เป็นเทคนิคที่มีประสิทธิภาพในการจำแนกประเภทของลูกค้าเช่นเดียวกัน เทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน แบบจำลองถูกสอนให้เรียนรู้จากชุดข้อมูลที่มีเลเบล เพื่อ

ทำนายผลจากตัวแปรที่ป้อนให้กับแบบจำลอง ยกตัวอย่างแบบจำลองที่ใช้เทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอนในการจำแนกประเภทของลูกค้า เช่น แบบจำลองต้นไม้ตัดสินใจ (Decision Tree) แบบจำลองDecision Tree เป็นแบบจำลองจำแนกประเภทข้อมูลที่ได้รับความนิยม เนื่องจากเป็นแบบจำลองที่ง่ายในการใช้งาน แบบจำลองDecision Tree ทำการจำแนกประเภทของข้อมูลโดยใช้วิธีการค่อยๆแบ่งข้อมูลออกเป็นกลุ่มย่อย โดยการแบ่งข้อมูลนี้จะอาศัยฟีเจอร์ที่มีความสำคัญเป็นเกณฑ์ในการแบ่ง ในงานประเภทของการจัดกลุ่มลูกค้า แบบจำลองDecision Tree ใช้การแตกกฎการตัดสินใจ เพื่อจำแนกลักษณะเฉพาะและพฤติกรรมของกลุ่มลูกค้าที่แตกต่างกัน ทำให้ธุรกิจเข้าใจข้อมูลเชิงลึกของลูกค้าที่มีคุณลักษณะ และรูปแบบที่แตกต่างกันในแต่ละกลุ่ม สิ่งนี้ช่วยให้ธุรกิจเข้าใจปัจจัยที่มีอิทธิพลต่อพฤติกรรม ความชื่นชอบ และการตัดสินใจซื้อของลูกค้าในแต่ละกลุ่มได้อย่างชัดเจน

ในงานวิจัยนี้ศึกษาเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอนด้วยแบบจำลองจัดกลุ่ม K-Means เพื่อจัดกลุ่มลูกค้าบัตรเครดิต โดยจัดกลุ่มจากลักษณะทางประชากรศาสตร์ และพฤติกรรมของลูกค้า และในลำดับถัดมาชุดข้อมูลลูกค้าบัตรเครดิตที่ถูกจัดกลุ่มนำมาทำนายกลุ่มด้วยแบบจำลอง Decision Tree ร่วมด้วยเพื่อให้การจัดกลุ่มมีประสิทธิภาพ และเพิ่มความน่าเชื่อถือให้แบบจำลอง โดยงานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนากลยุทธการจัดกลุ่มลูกค้า เพื่อให้ธุรกิจสามารถสร้างแคมเปญการตลาดออกแบบรายการส่งเสริมการขาย และนำเสนอโปรโมชั่นที่เหมาะสมกับความต้องการของลูกค้าในแต่ละกลุ่ม

งานวิจัยที่เกี่ยวข้อง

การแบ่งกลุ่มลูกค้า (Customer Segmentation) เป็นขั้นตอนหนึ่งของการจัดการการตลาด โดยมีทฤษฎี STP Theory [1] ประกอบไปด้วย Segmentation, Targeting และSegmentation โดยในงานวิจัยนำเสนอ Segmentation ซึ่งธุรกิจแบ่งลูกค้าออกเป็นกลุ่มย่อย โดยในแต่ละกลุ่มมีลักษณะเฉพาะที่แตกต่างกัน สามารถแบ่งลักษณะเฉพาะออกได้เป็น 4 ประเภทคือ

- แบ่งตามประชากรศาสตร์
- แบ่งตามพื้นที่
- แบ่งตามหลักจิตวิทยา
- แบ่งตามพฤติกรรม

เทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอนเป็นเครื่องมือที่มีประสิทธิภาพในการจัดกลุ่มลูกค้า ช่วยให้ธุรกิจสามารถจัดกลุ่มลูกค้าที่มีพฤติกรรมและลักษณะเฉพาะที่คล้ายกันอยู่ในกลุ่มเดียวกัน เนื่องด้วยมีเทคนิคในการจัดกลุ่มหลายรูปแบบ แต่มี 3 เทคนิคที่ได้รับความนิยมในการจัดกลุ่มลูกค้า ได้แก่ การจัดกลุ่ม K-Means, การจัดกลุ่มแบบมีลำดับชั้น (Hierarchical Clustering) และการจัดกลุ่มจากการกระจุกตัวของข้อมูล (Density-based Clustering หรือ DBSCAN) ในงานวิจัยที่ผ่านมาได้แสดงให้เห็นถึงประสิทธิภาพของการใช้เทคนิคการจัดกลุ่มเพื่อการแบ่งกลุ่มลูกค้าในอุตสาหกรรมต่างๆ งานวิจัย[2] ได้มีการศึกษาการนำเทคนิคการจัดกลุ่ม 3 แบบมาเปรียบเทียบกัน โดยใช้ในการแบ่งกลุ่มชุดข้อมูลลูกค้าในภาคธนาคาร ผลการศึกษาพบว่า การนำ DBSCAN

ร่วมกับ K-means สามารถดึงข้อมูลเชิงลึกของลูกค้ากลุ่มมูลค่าสูงได้ดี งานวิจัย [4] ศึกษาการแบ่งกลุ่มลูกค้าโดยใช้ชุดข้อมูลการขายออนไลน์ ในงานวิจัยนี้พบว่า DBSCAN สามารถแยกกลุ่มข้อมูลรบกวน (Noise) ออกจากชุดข้อมูลได้อย่างชัดเจน และนำไปสู่การพบกลุ่มที่มีมูลค่าสูงกับธุรกิจ จากการวิเคราะห์พฤติกรรมของลูกค้าในชุดข้อมูลการขายออนไลน์ ใช้วิธีการจัดกลุ่มด้วยแบบจำลอง [5] K-Means เพื่อระบุรูปแบบ และหาข้อมูลเชิงลึกว่าสินค้าใดทำยอดขายสูงสดให้กับธุรกิจ งานวิจัย [6] วิเคราะห์ชุดข้อมูลลูกค้าของโรงแรม ซึ่งประกอบไปด้วยพฤติกรรมของลูกค้า สถานที่ตั้ง และข้อมูลประชากรศาสตร์ โดยใช้อัลกอริทึมหลายรูปแบบในการจัดกลุ่มลูกค้า งานวิจัย [3] ศึกษาชุดข้อมูลลูกค้าของบริษัทบัตรเครดิตโดยใช้การจัดกลุ่มแบบ K-means และแบบลำดับชั้น (Hierarchical clustering) ผลการวิจัยสรุปว่า K-means เหมาะสมกว่าสำหรับชุดข้อมูลนี้ นอกจากนี้ยังมีการนำการ Principal Component Analysis ในการลดมิติของข้อมูล เพื่อปรับปรุงประสิทธิภาพของการจัดกลุ่ม

การใช้เทคนิคการจัดกลุ่มลูกค้าด้วยแบบจำลอง K-Means สิ่งที่สำคัญคือการกำหนดจำนวนกลุ่มที่เหมาะสม ในบางงานวิจัยใช้วิธีการหลายรูปแบบ เพื่อหาจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มลูกค้า ตัวอย่างเช่น การใช้วิธีการ Elbow method ร่วมกับ Silhouette score [4,6,3] หรือ ในบางงานวิจัยใช้ทั้ง Elbow method, Silhouette score และ The Davies–Bouldin index ในการวิเคราะห์ร่วมกัน[5,6,3]

โดยการผสมผสานเทคนิคการจัดกลุ่มร่วมกับเทคนิคการจำแนกประเภท ช่วยให้ธุรกิจสามารถตัดสินใจได้ดียิ่งขึ้น งานวิจัย [7] นำเทคนิคการจัดกลุ่ม และการจำแนกประเภทมาใช้กับชุดข้อมูลการศึกษา ผลการวิจัยพบว่าการจัดกลุ่มแบบ K-means สามารถจัดกลุ่มนักเรียนตามพฤติกรรมการใช้งานระบบ e-learning และใช้แบบจำลอง Decision Tree ในการวิเคราะห์ลักษณะเฉพาะของนักเรียนในแต่ละกลุ่ม งานวิจัย [8] นำเสนอการจัดการลูกค้าสัมพันธ์ โดยใช้การจัดกลุ่มด้วยแบบจำลอง K-Means และและการจำแนกประเภทด้วยแบบจำลอง Decision Tree เพื่อวิเคราะห์พฤติกรรมของลูกค้าของบริษัทโทรคมนาคม

ตารางที่ 1 แสดงถึงแอททริบิวต์ของชุดข้อมูล

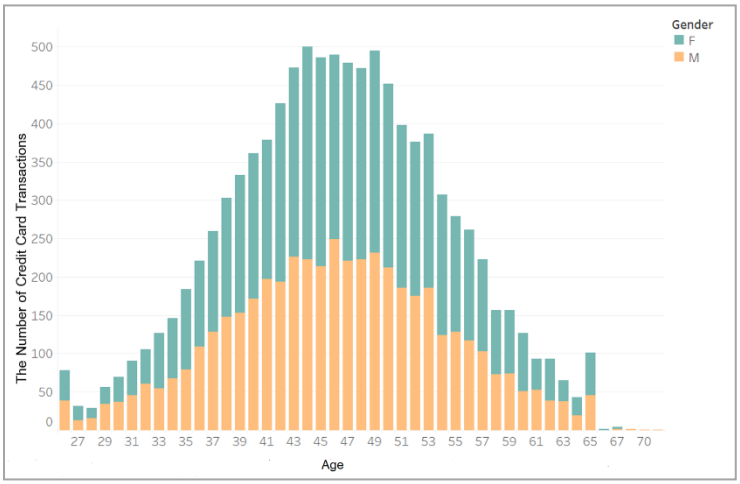
ลำดับ	ชื่อแอททริบิวต์	ข้อมูลภายในแอททริบิวต์	คำอธิบาย
1	Age	ปี	อายุของลูกค้า
2	Gender	M/F	เพศของลูกค้า
3	Education_Level	High School/Graduate/ College/Post-Graduate/ Doctorate/Uneducated/ Unknown	ระดับการศึกษาของลูกค้า
4	Marital_Status	Single/Married/Divorced/ Unknown	สถานภาพของลูกค้า
5	Months_on_book	เดือน	ระยะเวลาที่ลูกค้าถือบัตรเครดิต

6	Credit_Limit	จำนวนเงิน	วงเงินบัตรเครดิต
7	Total_Trans_Amount	จำนวนเงิน	ยอดเงินที่ใช้จ่ายบัตรเครดิต
8	Total_Trans_Count	จำนวนครั้ง	จำนวนครั้งของการใช้จ่ายบัตรเครดิต
9	Minimum_income	จำนวนเงิน	รายได้ขั้นต่ำที่ลูกค้าได้รับ
10	Max_income	จำนวนเงิน	รายได้สูงสุดที่ลูกค้าได้รับ

วิธีดำเนินการ

ขั้นตอนที่ 1 : ทำความเข้าใจข้อมูล และการเตรียมข้อมูล

การสำรวจข้อมูลเบื้องต้นเป็นขั้นตอนที่สำคัญในการทำความเข้าใจชุดข้อมูลลูกค้าบัตรเครดิต จากการวิเคราะห์ข้อมูลตัวอย่างจำนวน 10,127 คน พบว่ามีลูกค้าผู้หญิง 5,358 คน และลูกค้าผู้ชาย 4,769 คน ข้อมูลแสดงการกระจายตัวของอายุลูกค้าบัตรเครดิตมีช่วงอายุอยู่ระหว่าง 26 ถึง 73 ปี กลุ่มลูกค้าส่วนใหญ่อยู่ในช่วงอายุ 41-52 ปี (ตามภาพประกอบที่ 1) เมื่อพิจารณาฟีเจอร์ Total_Trans_Amt และTotal_Trans_Count จะเห็นว่ามีการสัมพันธ์กันระหว่างจำนวนธุรกรรมบัตรเครดิตกับยอดใช้จ่ายบัตรเครดิต จำนวนครั้งในการใช้บัตรสูงส่งผลต่อยอดใช้จ่ายบัตรเครดิตที่สูงขึ้น แสดงความสัมพันธ์เชิงบวกของจำนวนธุรกรรมกับยอดใช้จ่าย นอกจากนี้ ยังพบว่าลูกค้าที่ใช้บัตรเครดิตบ่อย และมีการใช้จ่ายสูงส่วนใหญ่เป็นลูกค้าผู้ชาย จากภาพประกอบที่ 2 แสดงให้เห็นถึงข้อมูลเชิงลึกเกี่ยวกับพฤติกรรมการใช้จ่ายมีความสัมพันธ์เชิงบวกกับการใช้บัตรเครดิตของลูกค้าผู้ชาย



ภาพประกอบที่ 1 แสดงการกระจายตัวของอายุของลูกค้า

ภาพประกอบที่ 2 แสดงความสัมพันธ์ระหว่างฟีเจอร์ Total_Trans_AmtและTotal_Trans_Count

เพื่อให้แบบจำลองได้เรียนรู้อย่างมีประสิทธิภาพ งานวิจัยนี้ได้ดำเนินการเตรียมข้อมูล (Data Preparation) โดยแบ่งเป็น 2 ขั้นตอน ดังนี้

- การแปลงค่าประเภทของข้อมูลให้เป็นตัวเลข (Encoding)

ชุดข้อมูลที่ใช้ในการวิจัยนี้มีฟีเจอร์แบบประเภท (Categorical Features) ได้แก่ Gender, Educational_Status และ Marital_Status ดังนั้น จึงต้องทำการแปลงค่าฟีเจอร์แบบประเภทเหล่านี้ให้เป็นตัวเลขด้วยการทำ Label Encoding

- การปรับช่วงค่าขอบเขตของฟีเจอร์ (Feature scaling)

ในงานวิจัยนี้ใช้วิธีการปรับช่วงค่าขอบเขตของฟีเจอร์ที่เป็นข้อมูลประเภทตัวเลขให้อยู่ในช่วงขอบเขตเดียวกันด้วยวิธี MinMaxScaler โดยเป็นการปรับค่าช่วงของข้อมูลให้อยู่ในช่วง [0,1]

ขั้นตอนที่ 2 : แบบจำลองแบบจัดกลุ่ม และการวัดประสิทธิภาพของแบบจำลองแบบจัดกลุ่ม

งานวิจัยนี้ใช้อัลกอริทึม K-means ในการจัดกลุ่มข้อมูล โดยการทดลองทั้งหมดใช้ไลบรารี scikit-learn ในภาษาไพธอน ในขั้นตอนแรกใช้เทคนิค K-means++ ในการสุ่มเพื่อหาตำแหน่งของจุดศูนย์กลาง (Centroids) หลังจากการทดลองแต่ละครั้ง ตำแหน่งของจุดศูนย์กลาง และผลลัพธ์การจัดกลุ่มจะถูกบันทึกไว้ และใช้เทคนิค Principal Component Analysis (PCA) ในการแสดงภาพ และวิเคราะห์ผล โดยการลดมิติของข้อมูลจาก 10 มิติเหลือ 2 มิติ

- Elbow Method

แบบจำลองK-Means เป็นแบบจำลองที่ต้องมีการกำหนดจำนวนกลุ่ม ซึ่ง Elbow method เป็นเทคนิคในการเลือกจำนวนกลุ่มที่เหมาะสม โดยใช้วิธีการลดค่าความคลาดเคลื่อน (Error) ซึ่งวัดโดยค่า Within-Cluster-Sum-of -Squares (WCSS) ซึ่งคำนวณจากผลรวมของระยะห่างระหว่างจุดข้อมูลและจุดศูนย์กลางของข้อมูลของทุกกลุ่ม เทคนิค Elbow method นิยมใช้ภาพประกอบในการพิจารณาจำนวนกลุ่มที่เหมาะสม โดยจุดที่เหมาะสมในการเลือกจำนวนกลุ่มคือ จุดที่มีการหักศอก ซึ่งที่เป็น

จุดที่ค่า Within-Cluster-Sum-of -Squared ลดลงน้อยเมื่อจำนวนกลุ่มเพิ่มขึ้น สูตรคำนวณ Within-Cluster-Sum-of -Squares ดังภาพประกอบ

$$WCSS = \sum_{C_k}^{C_m}\left(\sum_{d_i \, in \, C_i}^{d_m} distance(d_i, C_k)^2\right) \quad (1)$$

โดยที่ C คือจุดศูนย์กลางของกลุ่ม และd คือจุดข้อมูลของแต่ละกลุ่ม

- Silhouette Analysis

Silhouette coefficient score วัดคุณภาพของกลุ่ม โดยพิจารณาจากการแยกตัวกันระหว่างกลุ่ม และความเหมือนของจุดข้อมูลในแต่ละกลุ่ม โดยคำนวณระยะห่างระหว่างกลุ่มเทียบกับระยะห่างระหว่างจุดข้อมูลในกลุ่มเดียวกัน โดยค่าจะอยู่ระหว่าง -1 ถึง 1 ค่า Silhouette score เท่ากับ 1 แสดงถึงลักษณะของกลุ่มที่แยกตัวห่างจากกลุ่มอื่นอย่างชัดเจน และจุดข้อมูลในกลุ่มมีการเกาะกลุ่มกัน ค่า Silhouette score เท่ากับ 0 แสดงถึงขอบเขตของแต่ละกลุ่มมีการทับซ้อนกัน ไม่แยกตัวห่างกัน และค่า Silhouette score เท่ากับ -1 แสดงถึงจุดข้อมูลอยู่ผิดกลุ่ม สูตรการคำนวณ Silhouette coefficient score ดังภาพประกอบ

$$S = \frac{b-a}{max(a,b)} \quad (2)$$

โดยที่ a แทนค่าเฉลี่ยของระยะห่างระหว่างตัวอย่างกับจุดข้อมูลอื่นๆทั้งหมดในกลุ่มเดียวกัน และb แทนค่าเฉลี่ยของระยะห่างระหว่างตัวอย่างกับจุดข้อมูลทั้งหมดในกลุ่มถัดไปที่อยู่ใกล้ที่สุด

- Davies-Bouldin Index

Davies-Bouldin score คำนวณค่าเฉลี่ยความเหมือนระหว่างกลุ่ม โดยค่าความเหมือนวัดจากการเปรียบเทียบระยะห่างระหว่างกลุ่มกับขนาดของกลุ่ม โดยมีค่าตั้งแต่ 0 ขึ้นไป ซึ่งค่า Davies-Bouldin score เข้าใกล้ 0 แสดงถึงแต่ละกลุ่มมีการแยกตัวห่างกันอย่างชัดเจน สูตรการคำนวณ Davies-Bouldin score ดังภาพประกอบ

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j} R_{ij} \quad (3)$$

โดยที่ K แทนกลุ่ม, Rij แทนอัตราส่วนระหว่างกลุ่ม i และกลุ่ม j และMax แทนค่าอัตราส่วนภายในกลุ่มที่สูงที่สุด

- Calinski-Harabasz Index

Calinski-Harabasz score หรือที่รู้จักในชื่อ Variance Ratio Criterion คำนวณการกระจายตัวของข้อมูลระหว่างCluster และการกระจายตัวของข้อมูลภายในCluster โดยมีค่าตั้งแต่ 0 ขึ้นไป ซึ่งค่า Davies-Bouldin score สูงแสดงถึงถึงการเกาะกลุ่มกันของจุดข้อมูลในCluster และแต่ละCluster แยกตัวห่างกันอย่างชัดเจน สูตรการคำนวณ Calinski-Harabasz score ดังภาพประกอบ

$$CH = \frac{\frac{BGSS}{K-1}}{\frac{WGSS}{N-K}} = \frac{BGSS}{WGSS} \times \frac{N-K}{K-1} \quad (4)$$

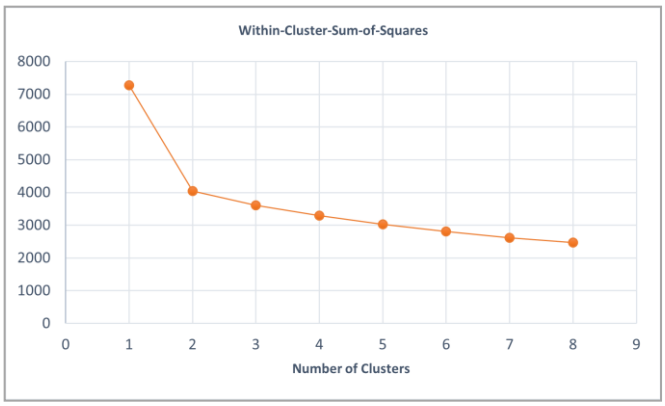
โดยที่ BGSS หรือ Between-Group Sum of Squares คำนวนระยะห่างระหว่างกลุ่มของแต่ละกลุ่ม, WGSS หรือ Within-Group Sum of Squares คำนวณระยะห่างของจุดข้อมูลภายในกลุ่มเดียวกัน, N แทนจำนวนจุดข้อมูลทั้งหมด และK แทนจำนวนกลุ่มทั้งหมด

ขั้นตอนที่ 3 : แบบจำลองจำแนกประเภท และการประเมินประสิทธิภาพของแบบจำลองจำแนกประเภท

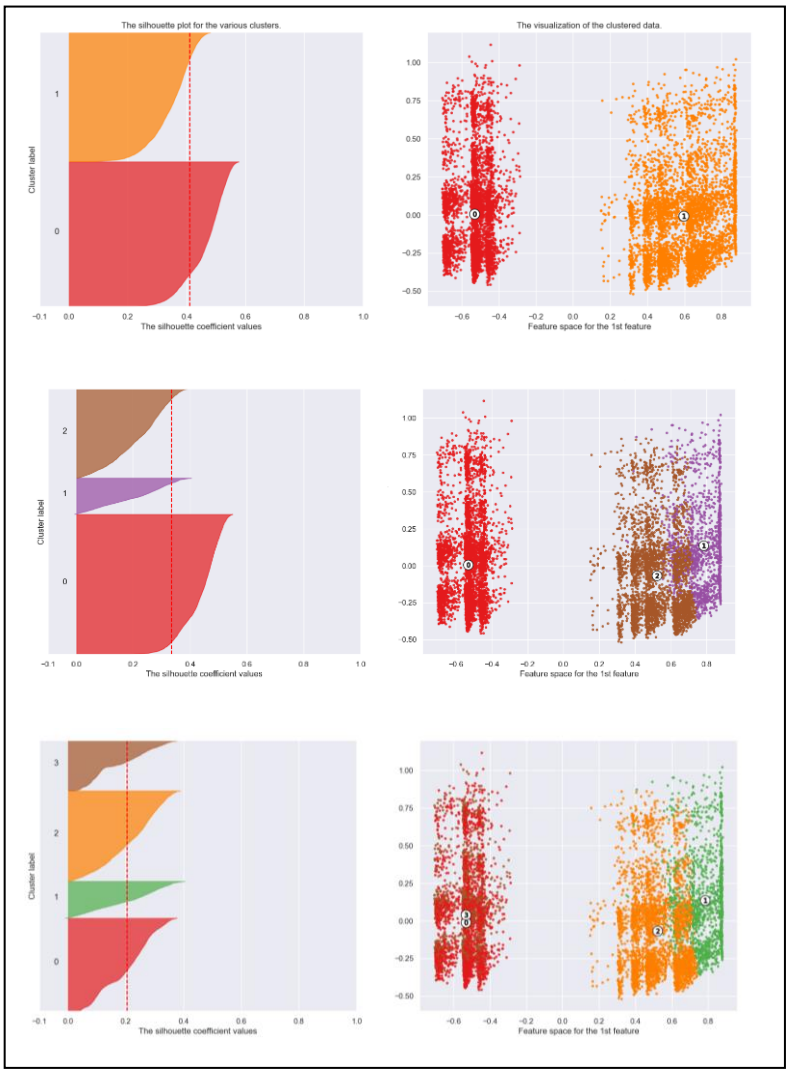
งานวิจัยนี้มีจุดมุ่งหมายเพื่อยกระดับประสิทธิภาพและความน่าเชื่อถือในการแบ่งกลุ่มลูกค้า โดยการทำนายกลุ่มลูกค้าด้วยแบบจำลองDecision Tree หลังจากแบ่งกลุ่มลูกค้าด้วยเทคนิคการจัดกลุ่มแล้ว นำชุดข้อมูลที่มีเลเบลมาใช้เป็นข้อมูลสำหรับแบบจำลองDecision Tree ในการทำนายกลุ่มข้อมูล โดยแบ่งชุดข้อมูลออกเป็นชุดข้อมูลสำหรับฝึกแบบจำลอง และชุดข้อมูลสำหรับทดสอบประสิทธิภาพแบบจำลองโมเดล โดยคำนึงถึงสัดส่วนของตัวอย่างในแต่ละกลุ่มให้มีความสม่ำเสมอ จากนั้นทำการทดลองแบบจำลอง Decision Tree รวมถึงการปรับจูนพารามิเตอร์ โดยงานวิจัยนี้ใช้เทคนิค GridSearchCV เพื่อหาค่าพารามิเตอร์ที่เหมาะสมจำนวน 5 รอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy, Precision, Recall, F1-Score

ผลการวิจัย

ในขั้นตอนการวิเคราะห์การจัดกลุ่ม เริ่มต้นด้วยการพิจารณาจำนวนกลุ่มที่เหมาะสม (k) โดยใช้ Elbow method จากภาพประกอบที่ 3 แสดงถึง ค่า WCSS สำหรับจำนวนกลุ่ม k เท่ากับ 2, 3 และ 4 ลดลงอย่างมีนัยสำคัญ จากนั้นคำนวณค่า Silhouette score สำหรับข้อมูลตัวอย่างในแต่ละกลุ่ม ในภาพประกอบที่ 4 แสดงค่า Silhouette score สำหรับข้อมูลตัวอย่างในแต่ละกลุ่มเมื่อกำหนดค่า k เท่ากับ 2, 3 และ 4 โดยเรียงลำดับคะแนนภายในแต่ละกลุ่มจากสูงไปต่ำ ขนาดของพื้นที่แสดงขนาดของแต่ละกลุ่ม เส้นประแสดงค่าเฉลี่ยของ Silhouette score ของแต่ละกลุ่ม

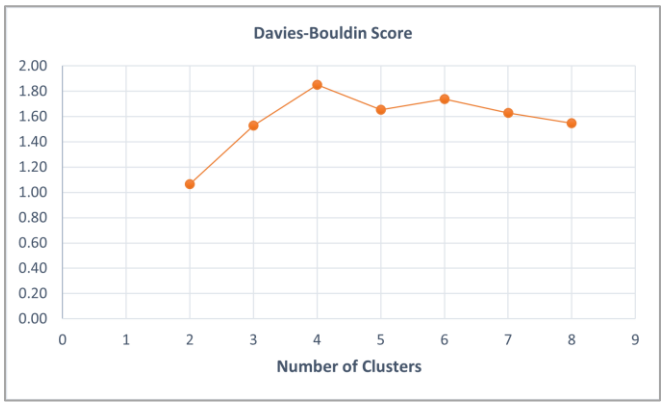


ภาพประกอบที่ 3 แสดงค่า Within-Cluster-Sum-of -Squared

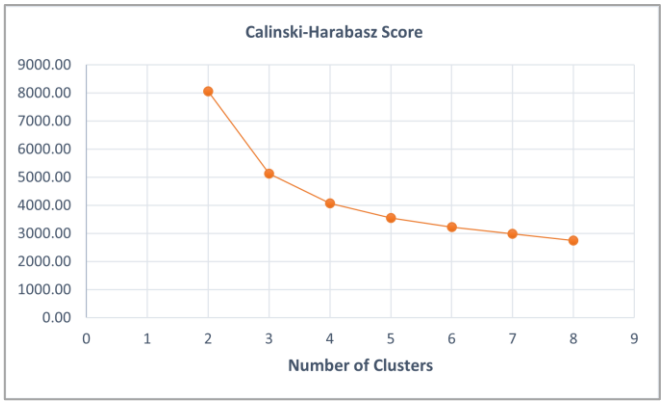


ภาพประกอบที่ 4 (ด้านซ้าย) แสดงค่าSilhouette scoreและค่าเฉลี่ยเมื่อ K=2,3 และ4 (ด้านขวา) แสดงจุดข้อมูลและCentroid (รูปวงกลม) ของแต่ละกลุ่ม

จากภาพประกอบที่ 4 (ด้านซ้าย) แสดงให้เห็นว่า การแบ่งกลุ่มเมื่อค่า k เท่ากับ 2 มีค่าเฉลี่ยของ Silhouette score สูงที่สุด คือ 0.4 รองลงมาคือการแบ่งกลุ่มเมื่อค่า k เท่ากับ 3 ซึ่งมีค่าเฉลี่ยเท่ากับ 0.33 และจากภาพประกอบที่ 4 (ด้านขวา) แสดงการกระจายของข้อมูลสองมิติหลังจากการใช้เทคนิค PCA ในการลดมิติข้อมูล กราฟแสดงการแบ่งกลุ่มข้อมูลอย่างชัดเจน โดยเฉพาะอย่างยิ่งเมื่อค่า k เท่ากับ 2 ซึ่งในกรณีที่ค่า k เท่ากับ 2 กลุ่มข้อมูลขนาดใหญ่แสดงด้วยสีแดง และกลุ่มข้อมูลขนาดเล็กแสดงด้วยสีส้ม และเมื่อค่า k เท่ากับ 3 กลุ่มข้อมูลขนาดเล็กกว่าได้แบ่งแยกออกเป็นอีกกลุ่มหนึ่ง



ภาพประกอบที่ 5 แสดงค่า Davies-Bouldin score



ภาพประกอบที่ 6 แสดงค่า Calinski-Harabasz score

ค่า Davies-Bouldin score ของแต่ละกลุ่มถูกคำนวณ และแสดงในภาพประกอบที่ 5 ผลลัพธ์ชี้ว่า การแบ่งกลุ่มแบบค่า k เท่ากับ 2 มีค่า DB score ต่ำที่สุด รองลงมาคือการแบ่งกลุ่มแบบค่า k เท่ากับ 3 และในภาพประกอบที่ 6 แสดงค่า Calinski-Harabasz score จากภาพแสดงถึงการแบ่งกลุ่มเมื่อค่า k เท่ากับ 2 มีค่า CB score สูงที่สุด รองลงมาคือการแบ่งกลุ่มเมื่อค่า k เท่ากับ 3

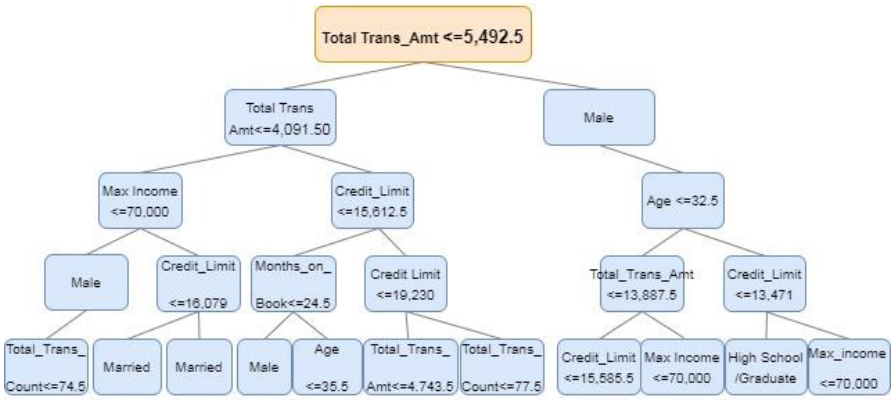
จากการประเมินเบื้องต้นนี้ แสดงให้เห็นว่า จำนวนกลุ่มที่เหมาะสมคือ ค่า k เท่ากับ 2 หรือ k เท่ากับ 3 อย่างไรก็ตาม ในงานวิจัยเลือกที่จะใช้การแบ่งกลุ่มแบบค่า k เท่ากับ 3 เนื่องจากเป็นงานวิจัยที่มีความมุ่งหมายที่จะนำไปใช้งานทางด้านการตลาด การ

แบ่งกลุ่มลูกค้าออกเป็นจำนวนหลายกลุ่ม ช่วยให้นักการตลาดทำความเข้าใจพฤติกรรมของลูกค้าได้อย่างเฉพาะเจาะจง และสามารถวางแผนกลยุทธการตลาดเฉพาะบุคคล เพื่อตอบสนองความต้องการของลูกค้าแต่ละกลุ่มได้อย่างมีประสิทธิภาพ

ตารางที่ 2 แสดงการวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy, Precision, Recall และF1-Score

Classification Model Performance	Metrics			
	Accuracy	*Precision*	*Recall*	*F1-Score*
Cluster 0	0.97	0.91	0.97	0.94
Cluster 1	1.00	1.00	1.00	1.00
Cluster 2	0.96	0.99	0.96	0.97
Average	0.98	0.97	0.98	0.97

ในขั้นตอนของแบบจำลองจำแนกประเภท Decision Tree โดยให้แบบจำลองเรียนรู้ชุดข้อมูลที่มีเลเบล ซึ่งแบบจำลองประสบความสำเร็จในการจำแนกประเภทข้อมูลทดสอบด้วยความแม่นยำ (Accuracy) ที่สูงถึง 0.98 นอกจากนี้ ผลการประเมินประสิทธิภาพของแบบจำลองได้ค่า Precision เท่ากับ 0.97 ค่า Recall เท่ากับ 0.98 และค่าF1-score เท่ากับ 0.97 ซึ่งบ่งชี้ถึงแบบจำลองทำงานมีประสิทธิภาพดีเยี่ยม



ภาพประกอบที่ 7 แสดงกฎการตัดสินใจของDecision Tree 4 ชั้น

เพื่อให้เข้าใจกระบวนการตัดสินใจของแบบจำลองDecision Tree มากขึ้น งานวิจัยนี้ได้ทำการวิเคราะห์ความสำคัญของฟีเจอร์ (Feature Importance) โดยวิเคราะห์ความสำคัญของฟีเจอร์ที่แบบจำลองใช้ในการเรียนรู้ ผลจากการวิเคราะห์แสดงให้เห็นว่าฟีเจอร์ Credit_Limit, Gender, Max_income และ Total_Trans_Amt มีอิทธิพลอย่างมากต่อกระบวนการตัดสินใจของ

แบบจำลอง จากภาพประกอบที่ 7 แสดงตัวอย่างกฎการตัดสินใจของDecision Tree 4 ชั้น ซึ่งผลลัพธ์ที่ได้สามารถนำไปใช้เป็นลักษณะเฉพาะของแต่ละกลุ่มลูกค้าที่สามารถอธิบายได้ และเป็นแนวทางสำหรับการแนะนำแคมเปญทางการตลาดต่อไป

ข้อเสนอแนะ

แนะนำแคมเปญทางการตลาด

ผลจากการวิเคราะห์การจัดกลุ่ม และลักษณะเฉพาะของกลุ่มลูกค้าที่ได้จากกฎการตัดสินใจของDecision Tree ผู้วิจัยขอเสนอแนวทางการตลาด เพื่อให้บริษัทบัตรเครดิตนำไปประยุกต์ใช้กับลูกค้าในแต่ละกลุ่มดังนี้

ลูกค้าในกลุ่มที่ 0 : Selective Spender

เป็นกลุ่มลูกค้าเพศชาย ส่วนใหญ่อยู่ในวัยกลางคนที่มีรายได้สูง ได้รับวงเงินบัตรเครดิตสูง มียอดใช้จ่ายบัตรเครดิตระดับน้อยถึงปานกลาง ถือบัตรเครดิตเป็นระยะเวลา 2-4 ปี และความถี่ในการใช้บัตรเครดิตอยู่ในระดับปานกลาง แคมเปญการตลาดสำหรับลูกค้ากลุ่ม Selective Spender

- กระตุ้นการใช้จ่ายบัตรเครดิตด้วยโปรแกรมสะสมคะแนนแลกรับรางวัล
- ร่วมรายการส่งเสริมการขายกับสินค้า หรือ บริการที่เหมาะสำหรับเพศชาย เพื่อกระตุ้นการใช้จ่ายผ่านบัตรเครดิต เช่น สินค้าประเภทเสื้อผ้า รองเท้า อุปกรณ์ไอที Gadget สินค้าที่เกี่ยวกับรถยนต์ บริการคาร์แคร์ บริการตัดผม เป็นต้น
- ร่วมรายการส่งเสริมการขายกับผลิตภัณฑ์ทางการเงิน และประกันชีวิต เนื่องจากช่วงอายุส่วนใหญ่ของลูกค้ากลุ่มนี้เป็นวัยกลางคนที่อยู่ในวัยที่ยังทำงานที่ต้องการสร้างความมั่นคง
- จัดกิจกรรมเพื่อสร้างความสัมพันธ์กับลูกค้ากลุ่ม Selective Spender เช่น กิจกรรมสอนตีกอล์ฟ แคมป์ปิ้ง พาชมการแข่งขันฟุตบอล เป็นต้น

ลูกค้าในกลุ่มที่ 1 : Moderate Spender

เป็นกลุ่มลูกค้าเพศหญิงที่มีรายได้น้อยถึงปานกลาง ได้รับวงเงินบัตรเครดิตน้อย มียอดใช้จ่ายบัตรเครดิตระดับน้อยจนถึงปานกลาง ถือบัตรเครดิตเป็นระยะเวลา 2-4 ปี และความถี่ในการใช้บัตรเครดิตอยู่ในระดับปานกลาง แคมเปญการตลาดสำหรับลูกค้ากลุ่ม Moderate Spender

- แคมเปญลดอัตราดอกเบี้ยเมื่อใช้จ่ายถึงยอดที่กำหนด
- โปรแกรม 0% สำหรับสินค้า หรือ บริการที่มีมูลค่าสูง
- กระตุ้นการใช้จ่ายบัตรเครดิตด้วยโปรแกรมสะสมคะแนนแลกรับของรางวัล
- เพิ่มวงเงินบัตรเครดิตให้กับลูกค้าที่ทีรายได้ระดับปานกลางขึ้นไป ที่มีประวัติการชำระหนี้ดี

- ร่วมรายการส่งเสริมการขายกับสินค้า หรือ บริการที่เหมาะสำหรับเพศหญิง เพื่อกระตุ้นการใช้จ่ายผ่านบัตรเครดิต เช่น ผลิตภัณฑ์ที่เกี่ยวกับความงาม สปา บริการเกี่ยวกับเส้นผม ทำเล็บ ผลิตภัณฑ์เกี่ยวกับแม่และเด็ก สินค้าประเภทเสื้อผ้า รองเท้า กระเป๋า เป็นต้น
- มอบบัตรกำนัลให้กับลูกค้าที่แนะนำเพื่อนมาใช้บัตรเครดิตผ่านโปรแกรมเพื่อนแนะนำเพื่อน
- จัดกิจกรรมเพื่อสร้างความสัมพันธ์กับลูกค้ากลุ่ม Moderate Spender เช่น กิจกรรมสอนแต่งหน้า ทำอาหาร หรือพาไปออกกำลังกาย เป็นต้น

ลูกค้าในกลุ่มที่ 2 : Strategic Spender

เป็นกลุ่มลูกค้าเพศชาย ส่วนใหญ่อยู่ในวัยกลางคนถึงวัยเกษียณมีรายได้ปานกลางถึงสูง ได้รับวงเงินบัตรเครดิตน้อยจนถึงปานกลาง มียอดใช้จ่ายบัตรเครดิตน้อยที่สุด ถือบัตรเครดิตเป็นระยะเวลา 2-4 ปี และความถี่ในการใช้บัตรเครดิตอยู่ในระดับปานกลาง แคมเปญการตลาดสำหรับลูกค้ากลุ่ม Strategic Spender

- แคมเปญลดอัตราดอกเบี้ยแบบขั้นบันได เมื่อใช้จ่ายถึงยอดที่กำหนด
- กระตุ้นการใช้จ่ายบัตรเครดิตด้วยโปรแกรมสะสมคะแนนแลกรับของรางวัล
- เพิ่มวงเงินบัตรเครดิตให้กับลูกค้าที่มีรายได้ระดับปานกลางขึ้นไป ที่มีประวัติการชำระหนี้ดี
- ร่วมรายการส่งเสริมการขายกับสินค้า หรือ บริการที่เหมาะสำหรับวัยผู้ใหญ่ เพื่อกระตุ้นการใช้จ่ายผ่านบัตรเครดิต เช่น ผลิตภัณฑ์ที่เกี่ยวกับการดูแลสุขภาพ วิตามิน อาหารเสริม แพ็คเกจตรวจสุขภาพกับโรงพยาบาล ผลิตภัณฑ์ที่เกี่ยวกับการออกกำลังกาย แพ็คเกจท่องเที่ยว เป็นต้น
- ร่วมรายการส่งเสริมการขายกับผลิตภัณฑ์ทางการเงิน และประกันชีวิต เนื่องจากในวัยนี้ต้องมีการจัดการทางการเงินเพื่อให้เพียงพอสำหรับการเกษียณ
- จัดกิจกรรมเพื่อสร้างความสัมพันธ์กับลูกค้ากลุ่ม Strategic Spender เช่น จัดทัวร์เดินป่าเพื่อสัมผัสธรรมชาติ จัดกิจกรรมออกกำลังกาย จัดกิจกรรมจิตอาสา และพาไปรับประทานอาหารร้านที่มีชื่อเสียง เป็นต้น

สรุปผลการวิจัย

งานวิจัยนี้แสดงให้เห็นว่า การผสมผสานเทคนิคการจัดกลุ่มแบบ K-Means ร่วมกับการจำแนกประเภทด้วยแบบจำลอง Decision Tree มีประสิทธิภาพในการแบ่งกลุ่มและทำนายพฤติกรรมของลูกค้า โดยใช้ฟีเจอร์ของชุดข้อมูล ในการวิเคราะห์การจัดกลุ่มสามารถแบ่งลูกค้าออกเป็น 3 กลุ่ม ซึ่งแบบจำลองDecision Tree นำผลลัพธ์การจัดกลุ่มมาใช้ในการจำแนกประเภทลูกค้าด้วยความแม่นยำ และมีประสิทธิภาพสูง นอกจากนี้แบบจำลอง Decision Tree ยังสามารถอธิบายผลลัพธ์ ซึ่งช่วยให้ธุรกิจสามารถทำความเข้าใจกระบวนการตัดสินใจเบื้องหลังการแบ่งกลุ่มลูกค้า ส่งผลให้ทีมการตลาดสามารถออกแบบแคมเปญ และโปรโมชั่นที่มีประสิทธิภาพ และตรงกับลักษณะเฉพาะของลูกค้าแต่ละกลุ่มได้อย่างแม่นยำ

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เอกสารอ้างอิง

[1] P. Kotler, K. Land Keller, “Marketing Management”, Pearson Education Inc., 2016, pp. 266-281

[2] M. Aliyev, E. Ahmador, H. Gadirli, “Segmenting Bank Customers via RFM Model and Unsupervised Machine Learning”, in arXiv.org, 2020

[3] A. Abdulhafedh, “Incorporating K-Means, Hierachical Clustering and PCA in Customer Segmentation”, in Journal of City and Development, 2021, Vol.3, No.1, 12-30.

[4] Mathes T., Sumathy G, Mahehwari A., “A Machine Learning approach to Segment the Customers of Online Sales Data for Better and Efficient Marketing Purposes”, in International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineer, IEEE.

[5] R. Punhani, V.P.s Arora, S. Sabitha, V.K. Shukla, “Application of Custering Algorithm for Effective Cluster Segmentation in E-Commerce”, in International Conference on Computational Intelligence and Knowledge Econimy, 2021, IEEE.

[6] D. Tesllenko, A. Sorokina, K. Smelyakov, O. Filipov, “Comparative Analysis of The Applicability of Five Clustering Algorithms for Market Segmentation”, in IEEE Open Conference of Electrical, Electronic and Information Sciences, 2023.

[7] S. Krizanic, “Educational Data Mining Using Cluster Analysis and Decision Tree Technique:A Case Study”, in International and Journal of Engineering Business Management, Vol. 12:1-9

[8] F. Abdi, A. Abolmakarem, “Customer Behavior Mining Framework (CBMF) Ussing Clustering and Classification Techniques”, in Journal of Industrial Engineering International (2019), 15 (Suppl 1): S1-S18