Evaluating Supervised Machine Learning Models

for Passenger Satisfaction Analysis

|  |  |
| --- | --- |
| |  | | --- | |  | |
| |  | | --- | |  | |  | |
| |  | | --- | | ปฏิภาณ เกริกชัยวัน | |
| |  | | --- | | บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ | | 2566 | |

|  |  |
| --- | --- |
| |  | | --- | |  | |
| |  | | --- | | ปฏิภาณ เกริกชัยวัน | |
| |  | | --- | | โครงร่างสารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร | | วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล | | คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ | | ปีการศึกษา 2566 | | ลิขสิทธิ์ของมหาวิทยาลัยศรีนครินทรวิโรฒ | |

|  |  |
| --- | --- |
| |  | | --- | |  | |
| |  | | --- | | PATIPAN KRIRKCHAIWAN | |
| |  | | --- | | A Master’s Project Proposal Submitted in Partial Fulfillment of the Requirements | | for the Degree of MASTER OF SCIENCE | | (Data Science) | | Faculty of Science, Srinakharinwirot University | | 2023 | | Copyright of Srinakharinwirot University | |

|  |  |
| --- | --- |
| สารนิพนธ์ | |
| เรื่อง | |
|  | |
| ของ | |
| ปฏิภาณ เกริกชัยวัน | |
|  | |
| ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล | |
| ของมหาวิทยาลัยศรีนครินทรวิโรฒ | |
|  | |
|  | |
| |  | | --- | |  | | (รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล) | | |
| คณบดีบัณฑิตวิทยาลัย | |
|  | |
| |  | | --- | |  | | |
|  | |
|  | |
| คณะกรรมการสอบปากเปล่าสารนิพนธ์ | |
|  | |
| .............................................. ที่ปรึกษาหลัก (อาจารย์ ดร.โสภณ มงคลลักษมี) | .............................................. ประธาน () |

**สารบัญ**

**หน้า**

สารบัญ จ

สารบัญตาราง ฉ

สารบัญรูปภาพ ช

บรรณานุกรม 2

ประวัติผู้เขียน 4

**สารบัญตาราง**

**หน้า**

No table of figures entries found.

**สารบัญรูปภาพ**

**หน้า**

No table of figures entries found.

**บทที่ 1**

**บทนำ**

* 1. **ที่มาและความสำคัญ**

ปัจจุบันการเดินทางโดยเครื่องบินเป็นการเดินทางที่สะดวกสบายและปลอดภัยที่สุด ประหยัดเวลาและค่าใช้จ่ายที่ถูกลงเนื่องจากปัจจุบันมีสายการบินเพิ่มขึ้นมากมายหลากหลายให้ผู้โดยสารเลือกใช้ การแข่งขันทางด้านธุรกิจการบินจึงสูง และนอกเหนือจากราคาแล้วการบริการ ความสะดวกสบาย ยังเป็นปัจจัยสำคัญต่อการเลือกใช้บริการของผู้โดยสารอีกด้วย สามารถสร้างความได้เปรียบทางด้านธุรกิจให้กับสายการบิน

* 1. **วัตถุประสงค์**
     1. เพื่อสร้างแบบจำลองที่มีประสิทธิภาพในทำนาย
     2. เพื่อสามารถหาสาเหตุของปัญหา
     3. เพื่อทดสอบหาแบบจำลองที่เหมาะสมกับข้อมูลที่สุด
  2. **วิธีการดำเนินการ**

1.3.1 ทำความเข้าใจกับข้อมูลความพึงพอใจของผู้โดยสาร โดยการวิเคราะห์ข้อมูลเชิงการตรวจสอบ (Exploratory Data Analysis) และทำความสะอาดข้อมูล(Data Claening)

1.3.2 ศึกษาแบบจำลองที่จะนำมาใช้กับข้อมูลความพึงพอใจของผู้โดยสาร และการใช้ค่าพารามิเตอร์ที่เหมาะสมกับชุดข้อมูล

1.3.3 ประเมินคุณภาพแบบจำลอง โดยพิจารณาจาก ค่าความถูกต้อง(Accuracy),ค่าความแม่นยำ(Precistion) ,ค่าความไว(Recall),F1-Score, ROC Graph and Area Under Curve (AUC)

1.3.4 สรุปผล

1.3.5 จัดทำบทความและวิทยานิพนธ์

* 1. **ขอบเขตและข้อจำกัดงานวิจัย**

ผู้วิจัยมีความสนใจถึงปัจจัยการให้บริการที่ส่งผลความพึงพอใจต่อผู้โดยสาร จึงใช้ข้อมูลจากเฉพาะสายการบินที่มีบริการเต็มรูปแบบ(Full Service)

โดยใช้U.S. Airline เป็น Publish Data Source เพื่อให้สามารถเทียบประสิทธิภาพกับงานวิจัยอื่นๆ

* 1. **ประโยชน์ที่คาดว่าจะได้รับ**1.5.1 ได้แบบจำลองที่มีประสิทธิภาพที่สุดในการทำนยความพึงพอใจของผู้โดยสาร

1.5.2 ได้รู้สาเหตุและปัจจัยที่ส่งผลต่อความพึงพอใจของผู้โดยสาร

**บทที่2**

**ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง**

ในบทนี้จะกล่าวถึง ทฤษฎีวรรณกรรมและงานวิจัยที่เกี่ยวข้อง ประกอบไปด้วย ความพึงพอใจของผู้โดยสาร ทฤษฏีแบบจำลองไบนาลี่คลาส ทฤษฏีการวัดประสิทธิภาพแบบจำลองแบบคลาสสิฟิเคชั่น

โดยมีรายละเอียดดังต่อไปนี้**2.1 ความพึงพอใจของผู้โดยสาร**

จากบทความวิจัยเรื่อง “A Critical Evaluation of Service Elements Related to Customer Satisfaction in the U.S. Airline Industry” จาก Dr.Adam F. Schvom กล่าวว่าความพึงพอใจของลูกค้าเป็นส่วนสำคัญในการดำเนินกิจการของสายการบินอย่างประสบความสำเร็จ ลูกค้าที่พึงพอใจมีแนวโน้มที่จะตัดสินใจใช้บริการเดินทางอีกครั้งและมีโอกาสแบ่งประสบการณ์ความพึงพอใจของพวกเขากับผู้อื่น ทำให้มีการใช้บริการของสายการบินเพิ่มขึ้น ในขณะที่สายการบินในสหรัฐฯหนึ่งรายกำลังมองหาวิธีใหม่ในการประสบความสำเร็จและชนะคู่แข่งในอุตสาหกรรมการบิน การระบุองค์ประกอบในการบริการที่เกี่ยวข้องกับความพึงพอใจของนักเดินทางทางอากาศและพื้นที่ที่ผู้ดดยสารไม่พึงพอใจคือสิ่งที่จำเป็น ในกระบวนการประเมินและแนะนำที่สำคัญ (CERP) นี้ ได้ทำการประเมินองค์ประกอบที่สำคัญ ที่ผู้โดยสารระบุว่าสำคัญสำหรับความพึงพอใจของพวกเขาในสายการบิน รวมถึงประเด็นที่สายการบินในสหรัฐฯหนึ่งสายไม่สามารถตรงตอบความคาดหวังของลูกค้าได้อย่างเพียงพอ ซึ่งจำเป็นต้องมีข้อเสนอแนะเพื่อปรับปรุงเพื่อเพิ่มความพึงพอใจของลูกค้า รายงานการประเมินและข้อแนะนำนี้จะถูกนำเสนอให้กับผู้บริหารของสายการบินใหญ่ในสหรัฐฯ โดยเมื่อมีหลักฐานธุรกิจที่มั่นใจสามารถสร้างขึ้นเพื่อเสนอการเปลี่ยนแปลงเพื่อเพิ่มระดับความพึงพอใจของลูกค้า สายการบินใหญ่ในสหรัฐฯจะดำเนินกิจกรรมเพื่อแก้ไขปัจจัยเสี่ยงต่อความพึงพอใจของผู้โดยสาร

**2.2 ทฤษฏีแบบจำลองไบนาลี่คลาสและทฤษฏีการวัดประสิทธิภาพแบบจำลอง**

การจัดประเภท(Classification) เป็นประเด็นสำคัญในการสำรวจและวิเคราะห์ข้อมูลขนาดใหญ่ การจัดประเภทนั้น เราใช้อัลกอริทึม (machine learning algorithm) เพื่อทดสอบคุณสมบัติของชุดข้อมูล เพื่อทำนายคุณลักษณะทางสถิติของชุดข้อมูล ซึ่งเรียกว่าคลาสลาเบล (class label) ซึ่งระบุคลาสเป้าหมาย กระบวนการนี้จะประเมินประสิทธิภาพของอัลกอริทึมแบบประเมินการทำนายอย่างแม่นยำ

**2.2.1 Logistic Regression with Binary Classification**

การวิเคราะห์การถดถอยโลจิสติก (Logistic regression analysis) เป็นการวิเคราะห์การถดถอยแบบ หนึ่งที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ (Qualitative variable) การวิเคราะห์การถดถอยโลจิสติกแบ่ง ออกเป็น 2 ประเภท คือ การวิเคราะห์การถดถอยโลจิ สติกแบบไบนารี (Binary logistic regression analysis) คือการวิเคราะห์การถดถอยที่ตัวแปรตามมี 2ค่าคือไม่ เกิดเหตุการณ์(y = 0) หรือเกิดเหตุการณ์(y = 1) การวิเคราะห์การถดถอยโลจิสติกมีวัตถุประสงค์เพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ จึงทำให้ถูกประยุกต์ใช้ในหลายสาขาวิชา เช่น ทางด้านการแพทย์ใช้สมการ โลจิสติก ทำนายโอกาสของการเกิดโรคต่างๆ ทางด้านธุรกิจใช้ ในการพยากรณ์ภาวะทางด้านเศรษฐกิจ เช่น โอกาสของการเกิดหนี้

เช่นเดียวทางด้านการบริการก็มีการประยุกต์ใช้การวิเคราะห์การถดถอยโลจิสติกใน

หลายๆ ด้าน เช่น การทำนายระดับความพึงพอใจของผู้บริโภค

เห็นได้ว่าการวิเคราะห์การถดถอยโลจิสติกสามารถนำมาประยุกต์ใช้ในหลายศาสตร์

โมเดลการวิเคราะห์การถดถอยคือ

A group of math equations

Description automatically generated

สำหรับการวิเคราะห์โลจิสติกที่เป็นแบบไบนารี ตัวแปรตาม (y) มีค่า คือ 0 และ 1 ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรทำนายจึงไม่อยู่ในรูปเชิงเส้น ซึ่งความสัมพันธ์ของตัวแปรตามและตัวแปรทำนายในการวิเคราะห์โลจิสติกจะอยู่ในรูปคล้ายตัว S

ดังรูปที่ 1

**A graph with blue dots

Description automatically generated**

**2.2.2 Decision Trees with classification**

Model Decision Tree เป็น Rule-Based Model ที่จะสร้างเงื่อนไข 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

โดย 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 การนำเอาผลรวมของค่าความน่าจะเป็นของเหตุการณ์ที่เราสนใจมาคูณด้วย (1 ลบ ค่าความน่าจะเป็นของเหตุการณ์ที่เราสนใจ)

A black and white math formula

Description automatically generated

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 ในเหตุการณ์ที่เราสนใจ คูณกับจำนวนข้อมูลของ Class ที่ i ในตัวแปรที่เราสนใจ และหารด้วยจำนวนข้อมูลทั้งหมดในตัวแปรที่เราสนใจ

A number and symbols on a white background

Description automatically generated

Entropy คือ คือ การวัด Randomness (ความไม่แน่นอน) ในการอธิบาย Target ของกลุ่มที่ถูกแบ่งออกมาจาก Feature นั่นหมายความว่าถ้าค่า Randomness ยิ่งต่ำก็ยิ่งแบ่งข้อมูลออกมาได้ดี

การคำนวณ Entropy

A black and white math symbol

Description automatically generated

ซึ่งสมการก็จะคล้ายคลึงกับค่า Gini Impurity เลย แต่ Entropy จะเป็นการคูณด้วย Log ของค่าความน่าจะเป็นของเหตุการณ์ที่เราสนใจ แทนการคูณด้วย 1 ลบ ค่าความน่าจะเป็นของเหตุการณ์ที่เราสนใจของ Gini Impurity ทำให้ค่า Entropy จะอยู่ในช่วงระหว่าง 0 – 1 แต่ Gini Impurity อยู่ในช่วงระหว่าง 0 – 0.5

**2.2.3 Random Forest**

**2.2.4 Support Vector Machines (SVM)**

**2.2.5 K-Nearest Neighbors (K-NN)**

**2.2.6 Naive Bayes**

**2.2.7 Neural Networks (Deep Learning)**

**2.2.8 Gradient Boosting Algorithms**

**บทที่3**

**วิธีการดำเนินงานวิจัย**

บรรณานุกรม

|  |
| --- |
| |  | | --- | | **บรรณานุกรม** | |  | |

ประวัติผู้เขียน

|  |
| --- |
| |  | | --- | | **ประวัติผู้เขียน** | |  | | |  |  | | --- | --- | | **ชื่อ-สกุล** | ปฏิภาณ เกริกชัยวัน | | **วัน เดือน ปี เกิด** | 17 กุมภาพันธ์ 2542 | | **สถานที่เกิด** | กรุงเทพมหานคร | | **ที่อยู่ปัจจุบัน** | 518/538 Supalai parkview ถ.รังสิต นครนายก ต.บึงยี่โถ อ.ธัญบุรี จ.ปทุมธานี 12130 | | |