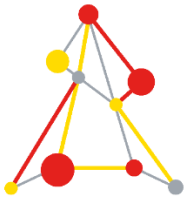


DSI@TU
Data Science & Innovation
Thammasat University

รายงานฉบับสมบูรณ์

การวิเคราะห์ข้อมูลสำหรับหาปัจจัยที่ทำให้มีโอกาสในการชนะของกีฬาเบสบอล

โดย

1.	นางสาวแพรวา ประสาทไทย	เลขทะเบียน 6424650031
2.	นายศิวกร นิตย์กิจสมบูรณ์	เลขทะเบียน 6424650395
3.	นายอธิชา เจริญธนกิจกุล	เลขทะเบียน 6424650494
4.	นางสาวเบญฤญา คำคงศักดิ์	เลขทะเบียน 6424650551
5.	นางสาววิชญาพร ธนสิทธิโชค	เลขทะเบียน 6424650627

อาจารย์ที่ปรึกษาโครงงาน: ผู้ช่วยศาสตราจารย์ ดร.ศรัณย์ กุลยานนท์

รายงานนี้เป็นส่วนหนึ่งของรายวิชา วข.314 โครงงานด้านการวิเคราะห์ข้อมูลสำหรับธุรกิจ

(DSI314 Business Analytics Capstone Project)

ภาคการศึกษา 1/2566

สารบัญ

สารบัญรูปภาพ

สารบัญตาราง

ตาราง **หน้า**

กิตติกรรมประกาศ

โครงงานฉบับนี้เป็นส่วนหนึ่งของวิชาวข. 314 โครงงานด้านการวิเคราะห์ข้อมูลสำหรับธุรกิจ โดยมีวัตถุประสงค์เพื่อให้คณะผู้จัดทำได้ฝึกการประยุกต์ใช้ความเชื่อมโยงกับวิชาต่าง ๆ ในโมดูล ได้แก่ วข. 310 การสำรวจและการเตรียมข้อมูล, วข.311 อัลกอริทึมของวิทยาศาสตร์ข้อมูล, วข.312 ระบบธุรกิจอัจฉริยะ และ วข.313 การวิเคราะห์การตลาด

ทั้งนี้โครงงานสามารถสำเร็จลุล่วงได้อย่างสมบูรณ์จาก ผู้ช่วยศาสตราจารย์ ดร.ศรัณย์ กุลยานนท์ ที่ได้สละเวลาอันมีค่าแก่คณะผู้จัดทำเพื่อให้คำปรึกษาและแนะนำ ตลอดจนตรวจทานแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่เป็นอย่างยิ่ง จนโครงงานฉบับนี้สำเร็จสมบูรณ์ลุล่วงได้ด้วยดี คณะผู้จัดทำขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

คณะผู้จัดทำ

บทสรุปผู้บริหาร

การแข่งขันกีฬาเบสบอล (Baseball) เป็นกีฬาที่มีความสนุกสนานและได้รับความนิยมสูง โดยเฉพาะในระดับเมเจอร์ลีก (MLB) ณ ประเทศอเมริกา โดยที่เป็นกีฬาที่ได้รับความสนใจจากแฟนคลับจำนวนมากและมีการติดตามอย่างใกล้ชิด ถึงแม้ว่าความนิยมของกีฬาเบสบอลในประเทศไทยอาจไม่มากเท่ากับประเทศอื่น ๆ แต่มีการจัดตั้งสมาคมกีฬาเบสบอลแห่งประเทศไทยและเข้าร่วมการแข่งขันระดับนานาชาติได้เป็นทางเลือกที่สร้างโอกาสให้กีฬาเบสบอลได้รับการยอมรับและพัฒนาอย่างต่อเนื่องในวงการกีฬาไทย

ทางคณะผู้จัดทำจึงสนใจและมีความต้องการวิเคราะห์กีฬาเบสบอล โดยใช้ข้อมูลสถิติการแข่งขันที่ทางเมเจอร์ลีกเบสบอล (MLB) ได้เก็บรวบรวมเอาไว้ เพื่อหาปัจจัยที่อาจส่งผลต่อโอกาสในการชนะ และนำไปต่อยอดในเชิงธุรกิจ เพื่อช่วยในการวางแผนกลยุทธ์ธุรกิจในการสร้างรายได้ เนื่องจากรายได้หลักขององค์กรมาจาก การที่ทีมชนะการแข่งขัน (Winner), การสนับสนุนจากสปอนเซอร์ (Sponsor), การขายบัตรเข้าชม (Ticket), การขายสินค้าที่เกี่ยวข้อง (Merchandise), และรายการทีวี (TV Show) ซึ่งการที่ทีมจะสามารถมีการสนับสนุนจากสปอนเซอร์ (Sponsor), การขายบัตรเข้าชม (Ticket), การขายสินค้าที่เกี่ยวข้อง (Merchandise), และรายการทีวี (TV Show) นั่นแปลว่า ทีมจะต้องเป็นทีมที่มีชื่อเสียง โดยการที่ทีมจะมีชื่อเสียงได้คือ การที่ทีมจะต้องชนะการแข่งขัน

กระบวนการวิเคราะห์กีฬาเบสบอล โดยใช้ข้อมูลสถิติการแข่งขันที่ทางเมเจอร์ลีกเบสบอล (MLB) ที่ได้เก็บรวบรวมเอาไว้ เพื่อหาปัจจัยที่อาจส่งผลต่อโอกาสในการชนะ ประกอบด้วยขั้นตอนดังนี้ ขั้นตอนที่ 1 การเก็บรวบรวมข้อมูลได้รวบรวมข้อมูลสถิติที่เกี่ยวข้องกับการแข่งขันในเมเจอร์ลีกเบสบอล (MLB) ทั้งทางทีมและนักกีฬาจากเว็บไซต์ https://baseballsavant.mlb.com/ เก็บรวบรวมด้วยวิธีการ Web Scraping และเก็บรวบรวมข้อมูลสถิติการแข่งขันในระยะเวลา 5 ปี โดยจะแบ่งเป็นฝั่ง Pitcher (ผู้ขว้าง) และ batter (ผู้ตี) ซึ่งมีจำนวนข้อมูล 1,047 และ 804 แถว ตามลำดับ ขั้นตอนที่ 2 คือ การทำ Exploratory Data Analysis (EDA) ด้วยการทำกราฟ Histogram และ Boxplot เพื่อดูการกระจายตัวของข้อมูล และทำกราฟ Box plot เพื่อตรวจสอบ Outliers ของแต่ละ Features ทั้งหมดในชุดข้อมูล Pitcher กับ Batter ขั้นตอนที่ 3 การทำ Data Preprocessing จะใช้เทคนิค RobustScaler เพื่อปรับสเกลข้อมูลที่ทนทานต่อ Outliers สามารถทำให้ข้อมูลปรับสเกลได้อย่างมีประสิทธิภาพโดยไม่ต้องได้รับผลกระทบจากค่า Outliers และใช้เทคนิค MinMaxScaler ช่วยปรับสเกลข้อมูลให้อยู่ในช่วง [0, 1] ซึ่งจะทำให้ปรับสเกลได้สมดุลและทำให้ค่าทุกค่าอยู่ในช่วงที่กำหนด และขั้นตอนที่ 4 คือการ Train and Evaluate model ในขั้นตอนนี้ได้มีการนำ Pipeline มาใช้สำหรับการ Training Model โดยใช้ LightGBM Regressor เป็นอัลกอริทึมที่ใช้ในการทำนายแบบ Regression เพื่อทำการสร้าง Model ของทั้ง Pitcher (ผู้ขว้าง) และ Batter (ผู้ตี) ซึ่งจากโมเดลจะได้ค่า Average Mean Absolute Error ของ Pitcher เฉลี่ยเท่ากับ 66.29 ถือเป็นค่า MAE ที่ต่ำ แสดงถึงความแม่นยำของโมเดลที่ดีในการทำนายข้อมูล และสำหรับข้อมูล

ของ Batter ได้ค่า Average Mean Absolute Error เฉลี่ยอยู่ที่ 121.28 ซึ่งเป็นผลลัพธ์ที่ทางคณะผู้จัดทำมองว่าค่อนข้างแย่ เนื่องจากเป็นค่าที่สูงและแสดงถึงประสิทธิภาพของโมเดลที่ไม่ดี และสุดท้ายทำการดึงค่าความสำคัญแต่ละตัวแปร (Feature Importance) จากการกำหนด Feature Selection จำนวน 10 ฟีเจอร์ในโมเดลของ Pitcher ได้แก่ PA, SO, 3B, BIP, OBP, Hits, BA, K%, BB และ HR ตามลำดับ และ 10 อันดับ Feature Importance ของ Batter ได้แแก่ PA ,HR, 3B, BIP, BA, OBP, HITS, SO, K% และ BB ตามลำดับ โดยเป้าหมายของเรา หรือ Pitches (จำนวนครั้งที่ตีโดน) ยิ่งสามารถตีโดนบอลไปมากเท่าไหร่ ก็จะทำให้มีโอกาสที่จะชนะมากขึ้น และ Feature Importance ยังส่งผลต่อโอกาสในการชนะที่จะสามารถตอบโจทย์ทางธุรกิจและสามารถสร้างกำไรให้แก่องค์กรได้มากขึ้น

บทที่ 1 บทนำ

1.1. ความสำคัญ / ที่มาของปัญหา

กีฬาเบสบอลเป็นหนึ่งในกีฬาที่ได้รับความนิยมทั่วโลกอย่างมาก โดยเฉพาะในเมเจอร์ลีกเบสบอล (MLB) ที่เป็นลีกยักษ์ใหญ่และมีแฟนคลับจำนวนมาก ถึงแม้ว่าในบางประเทศกีฬาเบสบอลไม่ได้รับความนิยมเท่ากับกีฬาอื่น แต่ก็ยังมีความสนใจและการติดตามในระดับท้องถิ่น ยกตัวอย่างเช่น ประเทศไทย ที่ถึงแม้ว่าจะมีการจัดตั้งสมาคมกีฬาเบสบอลแห่งประเทศไทยและมีการแข่งขัน เช่น การแข่งขันซีเกมส์, เอเชียนเกมส์ เป็นต้น แต่ก็ไม่ได้รับความนิยมในสังคมไทยมากเท่าไหร่

การแข่งขันในเมเจอร์ลีกเบสบอล (MLB) เป็นที่รู้จักในวงการกีฬาทั่วโลก โดยเป็นลีกยักษ์ใหญ่ที่มีประวัติความเป็นมายาวนาน บรรยากาศที่เต็มไปด้วยความเข้มข้นและความสนุกสนานทุกครั้งที่ทีมต่าง ๆ ที่เข้าร่วมการแข่งขันบุกเบิกสนาม เป็นที่นำมาซึ่งสาเหตุที่กีฬาเบสบอลกลายเป็นที่นิยมและเติบโตอย่างต่อเนื่อง

ทางคณะผู้จัดทำมีความสนใจในกีฬาเบสบอลเป็นอย่างมาก เนื่องจากเป็นกีฬาที่ได้รับความนิยม จึงอยากทำการศึกษาและวิเคราะห์ โดยจะใช้ข้อมูลสถิติการแข่งขันต่าง ๆ ที่ทาง MLB ได้รวบรวมเอาไว้ เพื่อนำข้อมูลที่ได้จากการวิเคราะห์ไปพัฒนาและเป็นประโยชน์ให้กับผู้ประกอบการหรือผู้ที่เกี่ยวข้องต่าง ๆ

1.2. วัตถุประสงค์ของโครงงาน

- ศึกษาและวิเคราะห์กีฬาเบสบอลที่มีความนิยมทั่วโลก โดยในที่นี้เน้นไปที่ข้อมูลสถิติการแข่งขันที่ทางเมเจอร์ลีกเบสบอล (MLB) ซึ่งเป็นลีกยักษ์ใหญ่ได้เก็บรวบรวมเอาไว้ เพื่อหาปัจจัยที่ส่งผลต่อโอกาสในการชนะ
- นำข้อมูลที่ได้จากการวิเคราะห์ไปใช้ประโยชน์กับผู้บริหารจัดการทีม เพื่อให้สามารถวางแผนจัดการทีมและนักกีฬาและสร้างกำไรได้เพิ่มขึ้น
- สามารถนำข้อมูลที่ได้จากการวิเคราะห์ไปใช้ประโยชน์ในการพัฒนาและปรับปรุงด้านทักษะของนักกีฬาเบสบอล ซึ่งมีประโยชน์ทั้งต่อนักกีฬาและผู้ที่เกี่ยวข้องต่าง ๆ ในสมาคมกีฬา
- ข้อมูลที่ได้สามารถทำให้เข้าใจและสนับสนุนในการพัฒนากีฬาเบสบอล และเป็นประโยชน์แก่ทุกคนที่สนใจและต้องการเรียนรู้เพิ่มเติมเกี่ยวกับกีฬาเบสบอล

1.3. ขอบเขตของโครงงาน

การรวบรวมข้อมูลและสถิติที่เกี่ยวข้องกับการแข่งขันในเมเจอร์ลีกเบสบอล (MLB) ของนักกีฬา เช่น ผลการแข่งขัน, สถิตินักกีฬาและข้อมูลอื่น ๆ ที่มีผลต่อผลการแข่งขัน มาวิเคราะห์เพื่อหาปัจจัยที่ส่งผลต่อโอกาสในการชนะ โดยใช้วิธีการดึงข้อมูลจากเว็บไซต์และใช้ Machine Learning ในกระบวนการวิเคราะห์ข้อมูลหาความสัมพันธ์และปัจจัยต่าง ๆ ที่มีผล เพื่อเป็นประโยชน์แก่ผู้ที่มีส่วนเกี่ยวข้องหรือผู้ที่สนใจ ทั้งในด้านของความเข้าใจ, การพัฒนาทักษะนักกีฬาและในด้านธุรกิจและการตลาด

1.4. ประโยชน์ที่ได้รับของโครงงาน

จากการวิเคราะห์ข้อมูลเพื่อหาปัจจัยที่มีผลต่อโอกาสในการชนะ ทำให้เกิดประโยชน์มากมายหลายด้าน ไม่ว่าจะเป็นในด้านนักกีฬาและทีม ที่จะได้ประโยชน์จากการนำไปพัฒนาทักษะและยกระดับประสิทธิภาพในการแข่งขัน หรือในด้านของธุรกิจที่ผู้บริหารจัดการทีมสามารถนำข้อมูลจาการวิเคราะห์ไปใช้ในการวางกลยุทธ์ การจัดทีมในการแข่งขัน หรือการวางแผนเพื่อพัฒนากีฬาและเพิ่มประสิทธิภาพในการจัดการกีฬา ทำให้เกิดประโยชน์สูงสุดทั้งในด้านธุรกิจและการตลาด รวมถึงเป็นประโยชน์ต่อผู้ที่มีความสนใจเกี่ยวกับกีฬาเบสบอลอีกด้วย

ดังนั้น โครงงานนี้จึงมีประโยชน์มากมายและสามารถสร้างคุณค่าต่อหลากหลายกลุ่มที่มีความเกี่ยวข้องหรือสนใจในกีฬาเบสบอล

บทที่ 2 ทบทวนวรรณกรรม

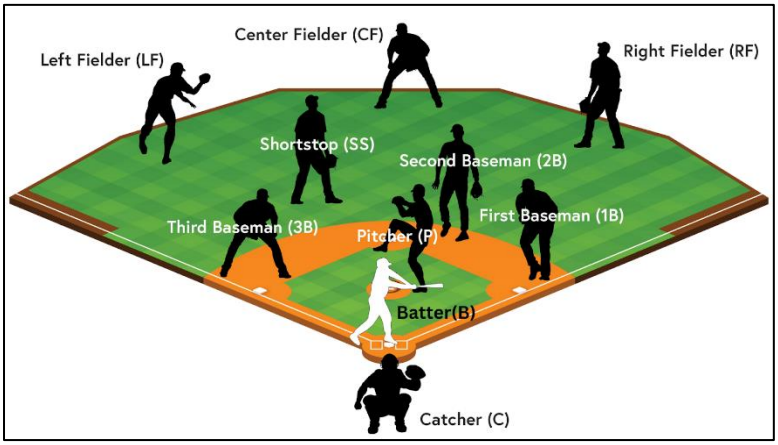
2.1. ความรู้พื้นฐาน

2.1.1. ความรู้ในเรื่องการเล่นเบสบอล

การเล่นในลีกเบสบอลมาจากสหรัฐอเมริกา (MLB - Major League Baseball) เป็นกีฬาเป้าหมายหนึ่งที่ได้รับความนิยมอย่างแพร่หลายทั่วโลก โดยมีกติกาและขั้นตอนการเล่น MLB ดังนี้

- การเล่นและกติกาพื้นฐาน: ในการแข่งขันใน MLB ประกอบไปด้วยการแข่งขัน 9 Innings โดยทีมที่ทำการทุจริตหรือมีสภาพแวดล้อมที่ไม่เหมาะสม การแข่งขันอาจถูกยุติก่อนสิ้นสุดเวลาที่กำหนด
- ผู้เล่น: ทีมประกอบด้วยผู้เล่น 9 คนที่เล่นฝ่ายโจมตีและฝ่ายป้องกัน และมีผู้เล่นสำรองสำหรับแต่ละตำแหน่ง
- โฮมเบส (Home Base) และเบสอื่น ๆ: มีโฮมเบส 3 อัน และเบสหลัก 4 อันในเกม การวิ่งรอบฐาน (Base Running) เป็นส่วนสำคัญของฝ่ายโจมตี
- การตีบอล (Batting): ผู้เล่นฝ่ายโจมตีจะต้องพยายามตีบอลที่โดนโยนจากผู้เสิร์ฟ และมุ่งหน้าสู่โฮมเบส
- การป้องกัน (Defense): ทีมป้องกันต้องพยายามหลีกเลี่ยงการให้คะแนนกับฝ่ายตรงข้าม โดยใช้กฎและการร่วมมือกันในทีม
- การส่งเสียบอล (Pitching): ผู้เสิร์ฟ (Pitcher) จะโยนบอลให้แก่ตัวต่อตัวฝ่ายตรงข้ามเพื่อให้เป็นลูกบอลที่ยากต่อการตี
- การใช้ Instant Replay: ในบางสถานการณ์ ผู้ตัดสินอาจใช้การตรวจสอบภาพในระบบ Instant Replay เพื่อตรวจสอบความถูกต้องของคำตัดสิน

2.1.2. ความสำคัญของ Pitcher และ Batter ในกีฬาเบสบอล



รูปที่ 2-1 อธิบายตำแหน่ง

Pitcher เป็นผู้เล่นที่มีบทบาทสำคัญในการควบคุมเกม ผู้เสิร์ฟจะต้องมีทักษะเฉพาะในการโยนบอลให้ทันตัวต่อตัวกับฝ่ายตรงข้าม และมีความยืดหยุ่นในการเปลี่ยนรูปแบบการโยนบอล เพื่อให้สร้างความสับสนและลูกบอลที่ยากต่อการตี ความสามารถในการควบคุมเกมและการทำกลยุทธ์ให้เกมเป็นไปตามที่ต้องการเป็นปัจจัยสำคัญที่ Pitcher ต้องคำนึงถึง

Batter คือตัวผู้เล่นที่ต้องการทำคะแนนโดยการตีบอลที่โยนมาจาก Pitcher การตีบอลต้องใช้ทักษะการอ่านการโยนบอลและการเรียนรู้ว่าควรตอบสนองอย่างไรให้เหมาะสม เพื่อทำให้มีโอกาสในการตีบอลไปไกล และการเตรียมความพร้อมในการต่อสู้กับท่าทางการโยนบอลที่หลากหลาย

2.1.3. ความรู้ในการสำรวจและการเตรียมข้อมูล

ในขั้นตอนการรวบรวมและการเตรียมข้อมูล คือกระบวนการที่จะทำให้ได้มาซึ่งข้อมูลที่ต้องการเพื่อตอบสนองตามวัตถุประสงค์ การเก็บรวบรวมข้อมูลมาจากแหล่งข้อมูลที่มีอยู่แล้ว ด้วยวิธีการ Web Scraping ซึ่งเป็นเทคนิคที่เหมาะสมในการดึงข้อมูลจากเว็บไซต์สาธารณะ โดยเลือกใช้ไลบารี BeautifulSoup และ Selenium มาใช้ประโยชน์ในการวิเคราะห์ข้อมูล เพื่อต่อยอดในวัตถุประสงค์ / ทำให้บรรลุวัตถุประสงค์ที่ตั้งไว้ได้ และหลังจากรวบรวมข้อมูลที่ต้องการได้แล้วจึงจะเริ่มทำความสะอาดข้อมูล เพื่อให้ข้อมูลพร้อมนำไปใช้งานในขั้นตอนถัดไป

2.2. งานที่เกี่ยวข้อง

2.2.1. Realtime Prediction of Match outcomes in Australian football

ผู้จัดทำ: Mitchell F Aarons, Chris M Young, Lyndell Bruce, Dan B Dwyer

วัตถุประสงค์: เพื่อทำนายผลการแข่งขันสุดท้ายของการแข่งขันฟุตบอลออสเตรเลียแบบเรียลไทม์ได้หรือไม่ นอกจากนี้โมเดลยังสามารถใช้เพื่อสร้างข้อมูลเชิงลึกเกี่ยวกับประสิทธิภาพของทีม และสนับสนุนการตัดสินใจของโค้ชในระหว่างการแข่งขัน

วิธีการ

- มีการใช้ฐานข้อมูลตัววัดประสิทธิภาพทางเทคนิคของทีม 168 รายการ จาก 829 รายการของการแข่งขัน Australian Football League ระหว่างปี 2017 ถึง 2021
- ใช้คุณลักษณะ (Feature) ทั้งสองชุด (Data-Driven และ Data-Informed) เพื่อฝึกและประเมินโมเดลทั้ง 6 ตัว (Generalized Linear Model, Random Forest และ Adaboost) ในการทำนายผลการแข่งขัน (ชนะ / แพ้) ภายใน 120 epochs (แสดงถึงเวลาที่ปรับ Normalization ในแต่ละการแข่งขัน)

ข้อดี:

- โมเดลทั้ง 6 มีประสิทธิภาพการทำงานที่ดี (ความแม่นยำเฉลี่ยอยู่ที่ 73.5-75.8%) ในการทำนายผลการแข่งขันเมื่อเทียบกับโมเดลเชิงสถิติที่ใช้คะแนนผลการแข่งขันเป็นหลัก (ความแม่นยำเฉลี่ยอยู่ที่ 77.4%)
- Data-Informed Feature Sets มีประสิทธิภาพดีกว่า Data-Driven
- ความแม่นยำในการทำนายเริ่มต้นของการแข่งขันอาจมีค่าต่ำ (45.7-48.8%) แต่เพิ่มขึ้นสูงสุดไปยังจุดสูงสุดใกล้สุดของการแข่งขัน (87.2-92.7%)

ข้อเสีย:

- การทำนายเริ่มต้นของการแข่งขันมีความแม่นยำที่ต่ำ
- ผลการวิจัยอาจมีความเชื่อมั่นที่ต่ำในช่วงเริ่มต้นของการแข่งขันแต่มีความเชื่อมั่นที่สูงในช่วงสิ้นสุดของการแข่งขัน

2.2.2. Learning from Machine Learning : Prediction of Age - Related athletic performance decline trajectories

ผู้จัดทำ: Christoph Hoog Antink, Anne K Braczynski, Bergita Ganse

วัตถุประสงค์: ทำนายการพัฒนาผลการแข่งขันของนักกีฬาจากการวัดเพียงครั้งเดียว โดยใช้เทคนิค Machine Learning เพื่อให้ได้ความแม่นยำและช่วยในการระบุปัจจัยที่ทำให้ประสิทธิภาพในการแข่งขันลดลงของนักกีฬาในแต่ละช่วงอายุ

วิธีการ

- การใช้ Machine Learning Approach นำเสนอการวิเคราะห์โดยใช้ Multilayer Neuronal Network เพื่อทำนายการพัฒนาผลการแข่งขันของนักกีฬาจากการวัดเพียงครั้งเดียว และเปรียบเทียบความแม่นยำกับโมเดลอื่น ๆ เช่น Average Decline Curve หรือ Individually Shifted Decline Curve
- การค้นพบปัจจัยที่กำหนดอัตราการลดลงของผลการแข่งขัน: ศึกษาเพิ่มเติมเกี่ยวกับปัจจัยที่มีผลต่อการลดลงของผลการแข่งขัน โดยการพิจารณาผลลัพธ์ของโมเดล

ข้อดี:

- ความแม่นยำในการทำนายที่สูง เทคโนโลยี Machine Learning สามารถทำนายผลการแข่งขันของนักกีฬาได้อย่างแม่นยำและน่าเชื่อถือได้มากขึ้นเมื่อเทียบกับโมเดลที่ใช้กันอย่างแพร่หลาย
- การระบุปัจจัยที่มีผลต่อการลดลงของผลการแข่งขัน การนำเสนอผลการวิเคราะห์ใหม่เกี่ยวกับปัจจัยที่มีผลต่อการลดลงของผลการแข่งขันจะช่วยในการเข้าใจเพิ่มเติมเกี่ยวกับกระแสการพัฒนาทางกีฬาของนักกีฬาในช่วงอายุ

ข้อเสีย:

- ความซับซ้อนของการนำเสนอผล การใช้เทคโนโลยี Machine Learning อาจมีความซับซ้อนในการทำนายและการตีความผลลัพธ์ที่จำเป็นต้องใช้ความเข้าใจในการวิเคราะห์และอธิบายได้อย่างถูกต้อง
- การจำกัดของข้อมูล การใช้ข้อมูลที่มีจำกัดอาจทำให้ความแม่นยำในการทำนายและการค้นพบปัจจัยมีข้อจำกัดได้

2.3. ตารางเปรียบเทียบโครงงานกับงานอื่นที่มีอยู่ในปัจจุบัน

ลักษณะสำคัญ (Feature)	Realtime Prediction of Match outcomes in Australian football	Learning from Machine Learning: Prediction of age-related athletic performance decline trajectories	โครงงานของคณะผู้จัดทำ
Data Analysis	/	/	/
Multilayer Neural Networks		/	
Adaboost	/		
Random Forest	/		
Generalized Linear Models	/		
Traditional linear		/	
Quadratic regression model		/	
มีการวางแผนกลยุทธ์การตลาด			/

ตารางที่ 2-1 ตารางเปรียบเทียบโครงงานกับงานอื่นที่มีอยู่ในปัจจุบัน

บทที่ 3 การดำเนินงานโครงงาน

3.1. การรวบรวมข้อมูล (Data Collection / Data Acquisition)

เป็นขั้นตอนของการรวบรวมข้อมูลที่เกี่ยวข้อง สำหรับนำไปสร้าง Model หรือสำหรับการวิเคราะห์ข้อมูลสถิติต่าง ๆ

3.1.1. แหล่งข้อมูลที่จะนำไปใช้เป็นเว็บไซต์ข้อมูลสถิติการแข่งขันต่าง ๆ ที่ทาง MLB ได้เก็บรวบรวมเอาไว้ ซึ่งมาจาก https://baseballsavant.mlb.com/ โดยจะทำการเก็บรวบรวมข้อมูลสถิติการแข่งขันในระยะเวลา 5 ปี ซึ่งจะมีปี 2019, 2020, 2021, 2022 และ 2023

3.1.2. เครื่องมือที่ใช้ในการทำ Web Scraping

คือ Selenium และ BeautifulSoup เนื่องจากเว็บไซต์มีการสร้างเนื้อหาแบบเปลี่ยนแปลงได้ (Dynamic Websites) เช่น เว็บไซต์ที่ใช้ JavaScript เพื่อโหลดข้อมูลหรือปรับแต่งหน้าเว็บไซต์ ดังนั้นจะใช้ Selenium เพื่อโหลดข้อมูลและ BeautifulSoup ใช้สำหรับการวิเคราะห์ HTML และดึงข้อมูลที่ต้องการหลังจากที่ข้อมูลถูกโหลด เพราะฉะนั้นการใช้ร่วมกันของ Selenium และ BeautifulSoup เป็นวิธีที่มีประสิทธิภาพในการทำ Web Scraping บนเว็บไซต์ที่มีเนื้อหาแบบไดนามิก

```
main_url = 'https://baseballsavant.mlb.com/'
minor_url = ['statcast_search?hfPT=&hfAB=&hfGT=R%7C&hfPR=&hfZ=&hfStadium=&hfBBL=&hfNewZones=&hfPull=&hfC=&hfSea=2023%7C2022%
             'statcast_search?hfPT=&hfAB=&hfGT=R%7C&hfPR=&hfZ=&hfStadium=&hfBBL=&hfNewZones=&hfPull=&hfC=&hfSea=2023%7C2022%

all_url = [main_url + url for url in minor_url]

def get_table_data(item):
    data = []
    table_body = item.find("div", class_="table-savant")
    rows = table_body.find_all('tr')
    for row in rows:
        headers = [header.text.strip() for header in row.find_all('th')]
        cols = row.find_all('td')
        cols = [ele.text.strip() for ele in cols]
        try:
            # Handle missing values
            for i in range(len(cols)):
                if cols[i] == "":
                    cols[i] = "-"
            # Combine headers and corresponding elements for each row
            rowwl = headers + cols
            # Filter out empty strings
            #rowwl = [ele for ele in rowwl if ele]
            data.append(rowwl)
        except:
            pass

    return data

all_df = []
sleep_time=2
for link in all_url:
    driver = webdriver.Chrome()
    driver.get(link)
    time.sleep(sleep_time)
    page_source = driver.page_source
    soup = BeautifulSoup(page_source, 'lxml')

    all_df.append(soup)


pitcher = pd.DataFrame(get_table_data(all_df[0]))
batter = pd.DataFrame(get_table_data(all_df[1]))
```

รูปที่ 3-1 Web Scraping

โดยข้อมูลของเราจะแบ่งเป็น Pitcher กับ Batter ซึ่ง Pitcher (ผู้ขว้าง) มีหน้าที่ขว้างลูกไปยัง Batter (ผู้รับ) แล้ว Batter จะมีหน้าที่ในการตีลูกที่ขว้างมาด้วยไม้เบสบอล และทำการเก็บข้อมูลทั้งหมดในรูปแบบของ Dataframe

ข้อมูลชุด Pitcher จะประกอบไปด้วย 1047 จำนวน

pitcher

Rk.	Player	Pitches	Total	Pitch %	PA	BIP	Hits	3B	HR	SO	...	Downward Movement w/ Gravity (in)	Glove/Arm-Side Movement (in)	Vertical Movement w/o Gravity (in)	Movement Toward/Away from Batter (in)	(MP
1	Cole, Gerrit RHP	14081	14089	99.9	3435	2073	653	5	120	1141	...	23.4	4.3	9.1	1.4	8
2	Nola, Aaron RHP	13553	13577	99.8	3456	2236	736	12	113	979	...	33.0	5.4	3.8	0.1	8
3	Castillo, Luis RHP	13152	13188	99.7	3252	2061	652	14	87	888	...	26.6	12.3	6.0	1.4	8
4	Giolito, Lucas RHP	12929	12941	99.9	3184	1994	657	18	124	905	...	23.6	5.4	12.3	0.6	8
5	Wheeler, Zack RHP	12925	12955	99.8	3320	2243	722	12	74	866	...	22.8	5.0	9.1	1.1	8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1043	Anderson, Drew RHP	505	506	99.8	128	98	30	0	4	17	...	28.0	3.0	8.5	1.7	8
1044	Gose, Anthony LHP	503	504	99.8	114	61	17	0	4	37	...	20.8	3.3	10.8	1.4	9
1045	Rodríguez, Manuel RHP	503	503	100.0	139	94	28	0	4	24	...	25.7	6.4	5.6	0.6	8
1046	Yamaguchi, Shun RHP	501	502	99.8	119	74	28	2	6	26	...	29.2	7.7	7.9	0.1	8
1047	Grotz, Zac RHP	501	501	100.0	118	73	25	0	4	22	...	32.1	8.6	4.1	3.2	8

1047 rows × 37 columns

รูปที่ 3-2 ข้อมูลชุด Pitcher

ข้อมูลชุด Batter จะประกอบไปด้วยข้อมูล 804 จำนวน

batter

Rk.	Player	Pitches	Total	Pitch %	PA	BIP	Hits	3B	HR	SO	...	Downward Movement w/ Gravity (in)	Glove/Arm-Side Movement (in)	Vertical Movement w/o Gravity (in)	Movement Toward/Away from Batter (in)	(MPH
1	Semien, Marcus	14081	14089	100.0	3180	2337	755	19	140	528	...	27.9	3.7	7.7	0.6	88
2	Goldschmidt, Paul	13553	13577	100.0	2919	1914	727	3	131	647	...	28.0	3.4	7.2	1.6	91
3	Olson, Matt	13152	13188	99.9	2839	1828	637	4	177	664	...	27.4	4.5	7.8	2.2	92
4	Freeman, Freddie	12929	12941	100.0	3029	2182	839	9	132	494	...	27.2	5.0	7.6	2.8	90
5	Soto, Juan	12925	12955	99.9	2816	1819	642	10	138	478	...	28.1	5.0	7.1	2.7	92
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
800	O'Grady, Brian	1020	1021	99.8	113	64	18	1	4	35	...	27.8	4.3	8.0	3.9	89
801	Williams, Nick	1019	1027	100.0	125	69	16	0	2	47	...	27.4	5.2	8.0	4.5	86
802	Butler, Lawrence	1019	1020	99.8	128	88	26	0	4	35	...	28.9	5.2	6.3	4.7	88
803	Fried, Max	1017	1019	97.0	117	75	26	0	0	33	...	26.1	5.9	8.7	2.9	88
804	Alcantara, Sandy	1016	1016	96.4	109	19	6	0	0	84	...	25.9	4.7	8.8	1.0	81

804 rows × 37 columns

รูปที่ 3-3 ข้อมูลชุด Batter

3.1.3. Features

โดยชุดข้อมูล Pitcher และ Batter มี Feature และจำนวนที่เหมือนกัน ซึ่งเท่ากับ 37 แต่รายละเอียดความหมายจะแตกต่างกันขึ้นอยู่กับบริบทของ Pitcher และ Batter

จำนวน Features ของ Pitcher มีรายละเอียด ดังนี้

1. Player = ผู้เล่นของ Pitcher
2. Pitches = จำนวนครั้งที่ขว้างแล้วทำให้ตีพลาด
3. Total = จำนวนครั้งที่ขว้างทั้งหมด
4. Pitch% = เปอร์เซ็นต์ที่ขว้างและทำให้ตีพลาดทั้งหมด
5. PA (Plate Appearances) = จำนวนครั้งที่ขว้างลูกแล้ว Batter ตีได้และวิ่งไปจุดต่อไปได้สำเร็จ
6. BIP (Balls In Play) = จำนวนครั้งที่ Pitcher สามารถขว้างลูกบอลให้เข้าเขตและเล่นต้อน Batter
7. Hits = จำนวนครั้งที่ขว้างลูกแล้วทำให้ Batter ตีได้โดยไม่มีข้อผิดพลาด
8. 3B (Triple) = จำนวนครั้งที่ขว้างลูกแล้ว Batter สามารถถึงฐานที่สามอย่างปลอดภัยโดยไม่มีข้อผิดพลาด
9. HR (Home runs) = จำนวนครั้งที่ขว้างลูกแล้วทำให้ Batter ได้ Home runs
10. SO (Strike out) = จำนวนครั้งที่ขว้างลูกแล้วทำให้ Batter ได้ Strike out หรือตีพลาดครบ 3 ครั้ง
11. K% = เปอร์เซ็นต์ที่ขว้างลูกแล้วทำให้ Batter ได้ Strike out หรือตีพลาดครบ 3 ครั้ง
12. BB (Bases on Balls) = จำนวนครั้งที่ขว้างลูกออกนอกเขตสี่เหลี่ยมและ Batter ไม่แกว่งไม้ตี
13. BB% = เปอร์เซ็นต์ที่ขว้างลูกออกนอกเขตสี่เหลี่ยมและ Batter ไม่แกว่งไม้ตี
14. BA (Batting Average) = ค่าเฉลี่ยที่ขว้างลูกแล้วทำให้ Batter ตีบอลและทำแต้มได้
15. xBA (Expected Batting Average) = ค่าความคาดหวังทางสถิติของค่าเฉลี่ยการขว้างโดยพิจารณาจากคุณภาพการสัมผัส โดยเน้นที่ความเร็วและมุมของลูกที่ขว้าง
16. OBP (On-Base Percentage) = จำนวนครั้งที่ขว้างลูกแล้วทำให้ Batter เข้าถึงฐานได้สำเร็จ โดยคำนึงถึงการตี และการเดิน
17. xOBP = เปอร์เซ็นต์การขว้างให้ Batter วิ่งไปยังฐานถัดไป
18. SLG (Slugging Percentage) = ใช้วัดความสามารถในการขว้างเป็นเปอร์เซ็นต์ โดยพิจารณาจากจำนวนฐานทั้งหมดที่ได้รับจากการตี
19. xSLG (Expected Slugging Percentage) = ค่าความคาดหวังของการขว้างโดยพิจารณาจากคุณภาพของการสัมผัส โดยคำนึงถึงปัจจัยต่าง ๆ เช่น ความเร็ว มุมการขว้าง และตัวชี้วัดอื่น ๆ ที่เกี่ยวข้อง เช่นเดียวกับสถิติอื่น ๆ ที่คาดการณ์ไว้ xSLG มีเป้าหมายที่จะวัดผลการปฏิบัติงานของผู้เล่นโดยพิจารณาจากคุณภาพของลูกที่ขว้างมากกว่าแค่ผลลัพธ์ที่แท้จริง

20. wOBA (Weighted On-Base Average) = ตัวชี้วัดเดียวที่ถ่วงน้ำหนักเพื่อวัดประสิทธิภาพการรับโดยรวมของผู้เล่น
21. xwOBA (Expected Weighted On-Base Average) = ใช้วิเคราะห์ขั้นสูงเพื่อประเมินประสิทธิภาพการเล่นเกมรับที่คาดหวังของผู้เล่นโดยพิจารณาจากคุณภาพการสัมผัสที่พวกเขาทำกับลูกคำนึงถึงปัจจัยต่าง ๆ รวมถึงความเร็ว มุมตี และคุณลักษณะของลูกตีอื่น ๆ เพื่อให้การประเมินความสามารถในการรับของผู้เล่นมีความเหมาะสมยิ่งขึ้น
22. Barrels = Pitcher ที่ขว้างได้ดีที่สุดทั้งองศาและความแรง
23. ISO (Isolated Power) = ความสามารถของ Pitcher ในการขว้างสำหรับฐานพิเศษ
24. Batter Run Value = ตัวชี้วัดที่ทำให้เกิดเหตุการณ์ เช่น การตี, การเดิน, Strikeouts, Home Runs และปัจจัยอื่น ๆ โดยกำหนดค่าการวิ่งให้กับแต่ละเหตุการณ์เหล่านี้ตามผลกระทบต่อการให้คะแนนการวิ่ง
25. Pitcher Run Value = ประสิทธิภาพโดยรวมของการวิ่งที่ป้องกันการทำคะแนนจาก Batter
26. Pitch (MPH) = ความเร็วของการขว้างลูก วัดด้วยหน่วย ไมล์ต่อชั่วโมง
27. Spin (RPM) = อัตราการหมุนของลูกที่ถูกขว้าง วัดด้วยหน่วย รอบต่อนาที
28. Downward Movement w/ Gravity (in) = การเคลื่อนที่ในแนวตั้งของลูกที่ขว้างโดยคำนึงถึงอิทธิพลของแรงโน้มถ่วง
29. Glove/Arm-Side Movement (in) = การเคลื่อนที่ในแนวนอนของลูกบอลที่ขว้างโดยสัมพันธ์กับด้านแขนของ Pitcher
30. Vertical Movement w/o Gravity (in) = การเคลื่อนที่ในแนวตั้งของลูกที่ขว้าง ไม่รวมอิทธิพลของแรงโน้มถ่วง
31. Movement Toward/Away from Batter (in) = การเคลื่อนที่ในแนวนอนของลูกที่ขว้างโดยสัมพันธ์กับ Batter
32. EV (MPH) = Exit Velocity ตัวที่ใช้วัดความเร็วของลูกหลังจากขว้าง วัดด้วยหน่วย ไมล์ต่อชั่วโมง
33. LA (°) = Launch Angle ตัวที่ใช้วัดมุมแนวตั้งที่ลูกออกจากมือหลังจากขว้างไป มีหน่วยเป็น องศา
34. Dist (ft) = ตัวที่ใช้วัดระยะทางที่ลูกตีมาจากจุดตี วัดด้วยหน่วย ฟุต
35. Hard Hit% = เปอร์เซ็นต์ของการขว้างลูกแล้วทำให้เกิดการตีในประเภท "ตีแรง" ลูกที่ตีแรง คือลูกที่ตีด้วยความเร็วที่สูงมาก ซึ่งบ่งบอกถึงแรงและกำลังที่สำคัญเบื้องหลังการสัมผัสลูก
36. Barrel/BBE% = เปอร์เซ็นต์ของการขว้างลูก Perfect ใช้เพื่ออธิบายลูกที่ตีด้วยความเร็วและมุมที่เหมาะสมที่สุด
37. Barrel/PA% = เปอร์เซ็นต์ของการขว้างและทำให้ Batter วิ่งไปจุดต่อไปได้สำเร็จ (PA) ใช้เพื่ออธิบายลูกที่ขว้างซึ่งมีการผสมผสานความเร็วและมุมที่เหมาะสมที่สุด

จำนวน Features ของ Batter มีรายละเอียด ดังนี้

1. Player = ผู้เล่นของ Batter
2. Pitches = จำนวนครั้งที่ตีโดน
3. Total = จำนวนครั้งที่ตีทั้งหมด
4. Pitch% = เปอร์เซ็นต์ที่ตีโดนจากทั้งหมด
5. PA (Plate Appearances) = จำนวนครั้งที่ตีได้และวิ่งไปจุดต่อไปได้สำเร็จ
6. BIP (Balls In Play) = จำนวนครั้งที่ Batter ตีลูกบอลได้ดีและสามารถสร้างโอกาสในการทำคะแนนให้กับทีมได้
7. Hits = จำนวนครั้งที่ตีได้และเข้าถึงฐานโดยไม่มีข้อผิดพลาด
8. 3B (Triple) = จำนวนครั้งที่ตีได้และสามารถไปถึงฐานที่สามได้อย่างไม่มีข้อผิดพลาด
9. HR (Home runs) = การตีลูกเลยออกนอกสนามจนผู้รับไม่สามารถรับลูกได้ซึ่งจะได้แต้มครบสี่ฐานทันที
10. SO (Strike out) = จำนวนครั้งที่ตี Strike out หรือตีพลาดครบ 3 ครั้ง
11. K% = เปอร์เซ็นต์ที่ได้ Strikeouts out หรือตีพลาดครบ 3 ครั้ง
12. BB (Bases on Balls) = Batter ไม่แกว่งไม้ตีเมื่อ Pitcher ขว้างออกนอกเขตตีสี่เหลี่ยม
13. BB% = เปอร์เซ็นต์ที่ Batter ไม่แกว่งไม้ตีเมื่อ Pitcher ขว้างออกนอกเขตตีสี่เหลี่ยม
14. BA (Batting Average) = ค่าเฉลี่ยที่ตีบอลได้สำเร็จและสามารถทำคะแนนได้
15. xBA (Expected Batting Average) = ค่าความคาดหวังทางสถิติของค่าเฉลี่ยการตีโดยพิจารณาจากคุณภาพการสัมผัส โดยเน้นที่ความเร็วและมุมของลูกที่ตี
16. OBP (On-Base Percentage) = จำนวนครั้งที่ Batter เข้าถึงฐานได้สำเร็จ โดยคำนึงถึงการตี และการเดิน
17. xOBP (Expected On-Base Percentage) = ค่าความคาดหวังที่ Batter จะเข้าถึงฐาน
18. SLG (Slugging Percentage) = ใช้วัดความสามารถในการตีด้วยพลังของ Batter เป็นเปอร์เซ็นต์ โดยพิจารณาจากจำนวนฐานทั้งหมดที่ได้รับจากการตี
19. xSLG (Expected Slugging Percentage) = ค่าความคาดหวังของการตีโดยพิจารณาจากคุณภาพของการสัมผัส โดยคำนึงถึงปัจจัยต่าง ๆ เช่น ความเร็ว มุมการตี และตัวชี้วัดอื่น ๆ ที่เกี่ยวข้อง เช่นเดียวกับสถิติอื่น ๆ ที่คาดการณ์ไว้ xSLG มีเป้าหมายที่จะวัดผลการปฏิบัติงานของผู้เล่นโดยพิจารณาจากคุณภาพของลูกที่ตีมากกว่าแค่ผลลัพธ์ที่แท้จริง
20. wOBA (Weighted On-Base Average) = ตัวชี้วัดเดียวที่ถ่วงน้ำหนักเพื่อวัดประสิทธิภาพการรุกโดยรวมของผู้เล่น

21. xwOBA (Expected Weighted On-Base Average) = ใช้วิเคราะห์ขั้นสูงเพื่อประเมินประสิทธิภาพการเล่นเกมรุกที่คาดหวังของผู้เล่น โดยพิจารณาจากคุณภาพการสัมผัสที่พวกเขาทำกับลูกคำนึงถึงปัจจัยต่างๆ รวมถึงความเร็ว มุมตี และคุณลักษณะของลูกตีอื่น ๆ เพื่อให้การประเมินความสามารถในการรุกของผู้เล่นมีความเหมาะสมยิ่งขึ้น
22. Barrels = Batter ที่ตีได้ดีที่สุดทั้งองศาและความแรง
23. ISO (Isolated Power) = ความสามารถของ Batter ในการตีสำหรับฐานพิเศษ
24. Batter Run Value = ตัวชี้วัดที่ทำให้เกิดเหตุการณ์เช่น การตี, การเดิน, Strikeouts, Home Runs และปัจจัยอื่น ๆ
25. Pitcher Run Value = ประสิทธิภาพโดยรวมการวิ่งที่ป้องกันการทำคะแนนจาก Pitcher
26. Pitch (MPH) = ความเร็วของการตีลูก วัดด้วยหน่วย ไมล์ต่อชั่วโมง
27. Spin (RPM) = อัตราการหมุนของลูกที่ถูกตี วัดด้วยหน่วย รอบต่อนาที
28. Downward Movement w/ Gravity (in) = การเคลื่อนที่ในแนวตั้งของลูกที่ตี โดยคำนึงถึงอิทธิพลของแรงโน้มถ่วง
29. Glove/Arm-Side Movement (in) = การเคลื่อนที่ในแนวนอนของลูกบอลที่ตี โดยสัมพันธ์กับด้านแขนของ Batter
30. Vertical Movement w/o Gravity (in) = การเคลื่อนที่ในแนวตั้งของลูกที่ตี ไม่รวมอิทธิพลของแรงโน้มถ่วง
31. Movement Toward/Away from Batter (in) = การเคลื่อนที่ในแนวนอนของลูกที่ตี
32. EV (MPH) = Exit Velocity ตัวที่ใช้วัดความเร็วของลูกเมื่อออกจากไม้ตีหลังจากสัมผัสกัน วัดด้วยหน่วยไมล์ต่อชั่วโมง
33. LA (°) = Launch Angle ตัวที่ใช้วัดมุมแนวตั้งที่ลูกออกจากไม้ตีหลังจากสัมผัสกันมีหน่วยเป็น องศา
34. Dist (ft) = ตัวที่ใช้วัดระยะทางที่ลูกตีมาจากจุดตีวัดด้วยหน่วย ฟุต
35. Hard Hit% = เปอร์เซ็นต์ของลูกที่ตีแล้วจัดอยู่ในประเภทตีแรงหรือลูกที่ตีด้วยความเร็วที่สูงมาก
36. Barrel/BBE% = เปอร์เซ็นต์ของการตีได้ลูก perfect ใช้เพื่ออธิบายลูกที่ตีด้วยความเร็วงและมุมที่เหมาะสมที่สุด
37. Barrel/PA% = เปอร์เซ็นต์ของการตีได้และวิ่งไปจุดต่อไปได้สำเร็จ (PA) ใช้เพื่ออธิบายลูกที่ตีซึ่งมีการผสมผสานความเร็วและมุมที่เหมาะสมที่สุด

3.2. การวิเคราะห์ภาพรวมข้อมูล (Exploratory Data Analysis / Data Visualization)

เป็นขั้นตอนที่สำคัญในการวิเคราะห์ข้อมูล เพื่อแสดงภาพรวมและทำให้เข้าใจข้อมูลมากขึ้น โดยจะมีรายละเอียด ดังนี้

3.2.1. Data Type

ตรวจสอบประเภทของข้อมูลในแต่ละชุดของ Pitcher และ Batter โดยใช้คำสั่ง info() ในการตรวจสอบ ซึ่งจะได้รายละเอียดของชุดข้อมูล Pitcher และ Batter ดังนี้

```
pitcher.info()

<class 'pandas.core.frame.DataFrame'>
Index: 1047 entries, 1 to 1047
Data columns (total 37 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Player                                 1047 non-null   object
 1   Pitches                                1047 non-null   object
 2   Total                                  1047 non-null   object
 3   Pitch %                                1047 non-null   object
 4   PA                                     1047 non-null   object
 5   BIP                                    1047 non-null   object
 6   Hits                                   1047 non-null   object
 7   3B                                     1047 non-null   object
 8   HR                                     1047 non-null   object
 9   SO                                     1047 non-null   object
 10  K%                                     1047 non-null   object
 11  BB                                     1047 non-null   object
 12  BB%                                    1047 non-null   object
 13  BA                                     1047 non-null   object
 14  xBA                                    1047 non-null   object
 15  OBP                                    1047 non-null   object
 16  xOBP                                   1047 non-null   object
 17  SLG                                    1047 non-null   object
 18  xSLG                                   1047 non-null   object
 19  wOBA                                   1047 non-null   object
 20  xwOBA                                  1047 non-null   object
 21  Barrels                                1047 non-null   object
 22  ISO                                    1047 non-null   object
 23  Batter Run Value                       1047 non-null   object
 24  Pitcher Run Value                      1047 non-null   object
 25  Pitch (MPH)                            1047 non-null   object
 26  Spin (RPM)                             1047 non-null   object
 27  Downward Movement w/ Gravity (in)      1047 non-null   object
 28  Glove/Arm-Side Movement (in)           1047 non-null   object
 29  Vertical Movement w/o Gravity (in)     1047 non-null   object
 30  Movement Toward/Away from Batter (in)  1047 non-null   object
 31  EV (MPH)                               1047 non-null   object
 32  LA (°)                                 1047 non-null   object
 33  Dist (ft)                              1047 non-null   object
 34  Hard Hit%                              1047 non-null   object
 35  Barrel/BBE%                            1047 non-null   object
 36  Barrel/PA%                             1047 non-null   object
dtypes: object(37)
```

รูปที่ 3-4 Data Type Pitcher

```
batter.info()

<class 'pandas.core.frame.DataFrame'>
Index: 804 entries, 1 to 804
Data columns (total 37 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Player                                 804 non-null    object
 1   Pitches                                804 non-null    object
 2   Total                                  804 non-null    object
 3   Pitch %                                804 non-null    object
 4   PA                                     804 non-null    object
 5   BIP                                    804 non-null    object
 6   Hits                                   804 non-null    object
 7   3B                                     804 non-null    object
 8   HR                                     804 non-null    object
 9   SO                                     804 non-null    object
 10  K%                                     804 non-null    object
 11  BB                                     804 non-null    object
 12  BB%                                    804 non-null    object
 13  BA                                     804 non-null    object
 14  xBA                                    804 non-null    object
 15  OBP                                    804 non-null    object
 16  xOBP                                   804 non-null    object
 17  SLG                                    804 non-null    object
 18  xSLG                                   804 non-null    object
 19  wOBA                                   804 non-null    object
 20  xwOBA                                  804 non-null    object
 21  Barrels                                804 non-null    object
 22  ISO                                    804 non-null    object
 23  Batter Run Value                       804 non-null    object
 24  Pitcher Run Value                      804 non-null    object
 25  Pitch (MPH)                            804 non-null    object
 26  Spin (RPM)                             804 non-null    object
 27  Downward Movement w/ Gravity (in)      804 non-null    object
 28  Glove/Arm-Side Movement (in)           804 non-null    object
 29  Vertical Movement w/o Gravity (in)     804 non-null    object
 30  Movement Toward/Away from Batter (in)  804 non-null    object
 31  EV (MPH)                               804 non-null    object
 32  LA (°)                                 804 non-null    object
 33  Dist (ft)                              804 non-null    object
 34  Hard Hit%                              804 non-null    object
 35  Barrel/BBE%                            804 non-null    object
 36  Barrel/PA%                             804 non-null    object
dtypes: object(37)
```

รูปที่ 3-5 Data Type Batter

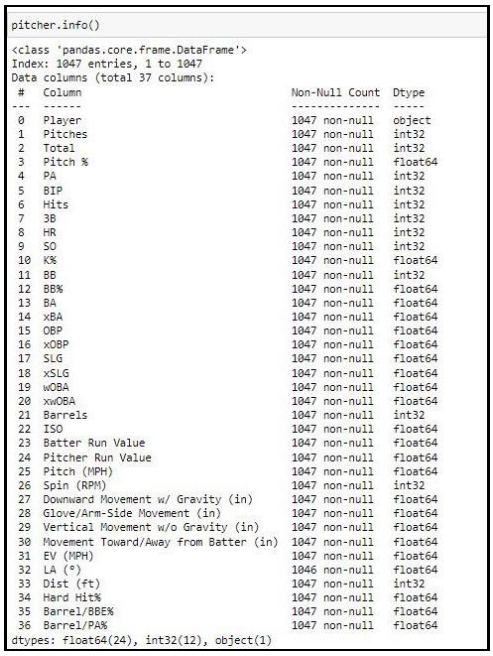
3.2.2. Convert Data Type

จากนั้นทำการแปลงประเภทของข้อมูลในคอลัมน์ทั้งหมดที่ไม่ถูกต้อง ยกเว้นคอลัมน์ที่ชื่อ "Player"

```
float_columns = ['Pitch %', 'K%', 'BB%', 'BA', 'xBA', 'OBP', 'xOBP', 'SLG', 'xSLG', 'wOBA', 'xwOBA', 'ISO', 'Batter Run Val
                 'Pitcher Run Value', 'Pitch (MPH)', 'Downward Movement w/\xa0Gravity (in)', 'Glove/Arm-Side Movement (in)',
                  'Vertical Movement w/o\xa0Gravity (in)', 'Movement Toward/Away from Batter (in)', 'EV (MPH)', 'LA (°)',
                 'Hard Hit%', 'Barrel/BBE%', 'Barrel/PA%']

int_columns = ['Pitches', 'Total', 'PA', 'BIP', 'Hits', '3B', 'HR', 'SO', 'BB', 'Barrels', 'Spin (RPM)', 'Dist (ft)']


# Pitcher

pitcher[float_columns] = pitcher[float_columns].replace('-', np.nan)
pitcher[float_columns] = pitcher[float_columns].astype(float)

pitcher[int_columns] = pitcher[int_columns].astype(int)


# Batter

#batter[float_columns] = batter[float_columns].replace('-', np.nan)
batter[float_columns] = batter[float_columns].astype(float)

batter[int_columns] = batter[int_columns].astype(int)
```

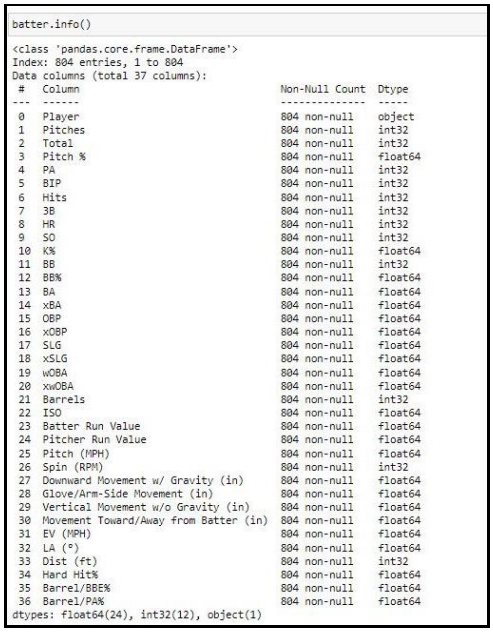
รูปที่ 3-6 Convert Data Type

โดยกำหนดให้ทุก Features ที่อยู่ใน float_columns ทำการแทนค่า “ – ” ด้วยค่า NaN สำหรับ Pitcher แล้วแปลงประเภทข้อมูลเป็น Float และ ในทุก Features ใน int_columns ทำการแปลงประเภทข้อมูลให้เป็น Integer ซึ่งจะได้ ดังนี้



```
pitcher.info()

<class 'pandas.core.frame.DataFrame'>
Index: 1047 entries, 1 to 1047
Data columns (total 37 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Player                                 1047 non-null   object
 1   Pitches                                1047 non-null   int32
 2   Total                                  1047 non-null   int32
 3   Pitch %                                1047 non-null   float64
 4   PA                                     1047 non-null   int32
 5   BIP                                    1047 non-null   int32
 6   Hits                                   1047 non-null   int32
 7   3B                                     1047 non-null   int32
 8   HR                                     1047 non-null   int32
 9   SO                                     1047 non-null   int32
 10  K%                                     1047 non-null   float64
 11  BB                                     1047 non-null   int32
 12  BB%                                    1047 non-null   float64
 13  BA                                     1047 non-null   float64
 14  xBA                                    1047 non-null   float64
 15  OBP                                    1047 non-null   float64
 16  xOBP                                   1047 non-null   float64
 17  SLG                                    1047 non-null   float64
 18  xSLG                                   1047 non-null   float64
 19  wOBA                                   1047 non-null   float64
 20  xwOBA                                  1047 non-null   float64
 21  Barrels                                1047 non-null   int32
 22  ISO                                    1047 non-null   float64
 23  Batter Run Value                       1047 non-null   float64
 24  Pitcher Run Value                      1047 non-null   float64
 25  Pitch (MPH)                            1047 non-null   float64
 26  Spin (RPM)                             1047 non-null   int32
 27  Downward Movement w/ Gravity (in)      1047 non-null   float64
 28  Glove/Arm-Side Movement (in)           1047 non-null   float64
 29  Vertical Movement w/o Gravity (in)     1047 non-null   float64
 30  Movement Toward/Away from Batter (in)  1047 non-null   float64
 31  EV (MPH)                               1047 non-null   float64
 32  LA (°)                                 1046 non-null   float64
 33  Dist (ft)                              1047 non-null   int32
 34  Hard Hit%                              1047 non-null   float64
 35  Barrel/BBE%                            1047 non-null   float64
 36  Barrel/PA%                             1047 non-null   float64
dtypes: float64(24), int32(12), object(1)
```

รูปที่ 3-7 Convert Data Type Pitcher



```
batter.info()

<class 'pandas.core.frame.DataFrame'>
Index: 804 entries, 1 to 804
Data columns (total 37 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Player                                 804 non-null    object
 1   Pitches                                804 non-null    int32
 2   Total                                  804 non-null    int32
 3   Pitch %                                804 non-null    float64
 4   PA                                     804 non-null    int32
 5   BIP                                    804 non-null    int32
 6   Hits                                   804 non-null    int32
 7   3B                                     804 non-null    int32
 8   HR                                     804 non-null    int32
 9   SO                                     804 non-null    int32
 10  K%                                     804 non-null    float64
 11  BB                                     804 non-null    int32
 12  BB%                                    804 non-null    float64
 13  BA                                     804 non-null    float64
 14  xBA                                    804 non-null    float64
 15  OBP                                    804 non-null    float64
 16  xOBP                                   804 non-null    float64
 17  SLG                                    804 non-null    float64
 18  xSLG                                   804 non-null    float64
 19  wOBA                                   804 non-null    float64
 20  xwOBA                                  804 non-null    float64
 21  Barrels                                804 non-null    int32
 22  ISO                                    804 non-null    float64
 23  Batter Run Value                       804 non-null    float64
 24  Pitcher Run Value                      804 non-null    float64
 25  Pitch (MPH)                            804 non-null    float64
 26  Spin (RPM)                             804 non-null    int32
 27  Downward Movement w/ Gravity (in)      804 non-null    float64
 28  Glove/Arm-Side Movement (in)           804 non-null    float64
 29  Vertical Movement w/o Gravity (in)     804 non-null    float64
 30  Movement Toward/Away from Batter (in)  804 non-null    float64
 31  EV (MPH)                               804 non-null    float64
 32  LA (°)                                 804 non-null    float64
 33  Dist (ft)                              804 non-null    int32
 34  Hard Hit%                              804 non-null    float64
 35  Barrel/BBE%                            804 non-null    float64
 36  Barrel/PA%                             804 non-null    float64
dtypes: float64(24), int32(12), object(1)
```

รูปที่ 3-8 Convert Data Type Batter

จากนั้นทำการแสดงผลสรุปชุดข้อมูลสถิติของ Pitcher และ Batter ตามลำดับ

0	count	mean	std	min	5%	25%	50%	75%	90%	95%	99%	max
Pitches	1,047.00	2,899.80	2,639.43	501.00	578.60	1,057.00	1,981.00	3,640.50	6,770.80	8,938.70	12,362.88	14,081.00
Total	1,047.00	2,905.69	2,645.82	501.00	579.30	1,058.00	1,984.00	3,648.00	6,779.60	8,954.10	12,387.02	14,089.00
Pitch %	1,047.00	99.80	0.14	98.90	99.60	99.70	99.80	99.90	100.00	100.00	100.00	100.00
PA	1,047.00	734.31	674.28	114.00	146.00	267.50	503.00	933.00	1,712.80	2,293.80	3,131.48	3,456.00
BIP	1,047.00	494.60	462.37	61.00	98.00	178.50	333.00	618.00	1,159.00	1,574.10	2,088.16	2,442.00
Hits	1,047.00	161.48	150.72	16.00	32.00	59.00	108.00	201.50	371.00	515.70	687.10	857.00
3B	1,047.00	2.80	3.15	0.00	0.00	1.00	2.00	4.00	7.00	10.00	14.00	21.00
HR	1,047.00	23.97	22.69	0.00	4.00	9.00	16.00	30.00	58.40	74.00	102.54	140.00
SO	1,047.00	170.71	167.87	15.00	28.00	57.00	110.00	220.50	398.80	537.00	821.70	1,141.00
K%	1,047.00	22.60	5.08	9.30	14.80	19.15	22.20	25.60	29.20	31.50	36.83	42.20
BB	1,047.00	60.39	49.92	4.00	13.00	25.00	44.00	77.00	129.00	167.00	239.16	287.00
BB%	1,047.00	9.05	2.74	2.40	5.20	7.10	8.70	10.60	12.64	14.07	16.90	22.60
BA	1,047.00	0.25	0.03	0.15	0.20	0.23	0.25	0.27	0.29	0.30	0.34	0.36
xBA	1,047.00	0.25	0.03	0.16	0.20	0.23	0.25	0.26	0.28	0.29	0.31	0.35
OBP	1,047.00	0.32	0.03	0.22	0.27	0.30	0.32	0.34	0.37	0.38	0.42	0.44
xOBP	1,047.00	0.32	0.03	0.22	0.28	0.30	0.32	0.34	0.36	0.38	0.40	0.43
SLG	1,047.00	0.42	0.07	0.19	0.32	0.37	0.41	0.46	0.51	0.54	0.62	0.79
xSLG	1,047.00	0.41	0.06	0.25	0.32	0.37	0.41	0.45	0.49	0.51	0.56	0.72
wOBA	1,047.00	0.32	0.04	0.23	0.27	0.30	0.32	0.34	0.37	0.39	0.42	0.48
xwOBA	1,047.00	0.32	0.03	0.23	0.28	0.30	0.32	0.34	0.36	0.38	0.40	0.46
Barrels	1,047.00	38.29	36.23	2.00	7.00	14.00	26.00	48.00	92.00	121.00	161.54	227.00
ISO	1,047.00	0.17	0.05	0.04	0.11	0.14	0.17	0.20	0.23	0.26	0.30	0.43
Batter Run Value	1,047.00	-1.01	21.79	-142.50	-37.48	-8.45	1.70	10.60	19.54	28.14	45.72	81.50
Pitcher Run Value	1,047.00	1.01	21.79	-81.50	-28.14	-10.60	-1.70	8.45	25.44	37.48	82.66	142.50
Pitch (MPH)	1,047.00	88.95	2.71	78.20	84.40	87.40	89.10	90.60	92.20	93.20	94.95	97.00
Spin (RPM)	1,047.00	2,244.11	187.09	1,596.00	1,949.10	2,121.50	2,239.00	2,368.50	2,486.00	2,547.70	2,719.62	2,837.00
Downward Movement w/ Gravity (in)	1,047.00	27.68	4.44	15.50	21.03	24.80	27.30	30.20	33.40	35.37	40.41	47.60
Glove/Arm-Side Movement (in)	1,047.00	5.01	3.27	0.00	0.50	2.40	4.60	7.25	9.54	10.87	13.76	17.10
Vertical Movement w/o Gravity (in)	1,047.00	7.53	3.10	-3.70	1.90	5.60	7.80	9.60	11.30	12.27	14.50	17.70
Movement Toward/Away from Batter (in)	1,047.00	1.53	1.45	0.00	0.10	0.50	1.10	2.00	3.70	4.77	6.35	8.20
EV (MPH)	1,047.00	88.80	1.46	83.70	86.23	87.90	88.90	89.80	90.60	91.10	91.95	93.50
LA (°)	1,046.00	12.96	5.01	-7.00	4.60	10.00	13.30	16.40	18.95	20.90	23.66	28.10
Dist (ft)	1,047.00	158.41	16.90	84.00	128.30	148.00	160.00	170.00	179.00	183.70	193.00	200.00
Hard Hit%	1,047.00	38.68	4.40	23.30	31.13	35.90	38.70	41.50	44.10	45.57	49.30	51.90
Barrel/BBE%	1,047.00	7.67	2.07	2.00	4.60	6.50	7.80	9.10	10.50	11.30	13.35	18.00
Barrel/PA%	1,047.00	5.28	1.50	1.40	3.10	4.30	5.20	6.20	7.20	7.90	9.71	12.30

รูปที่ 3-9 ข้อมูลสถิติ Pitcher

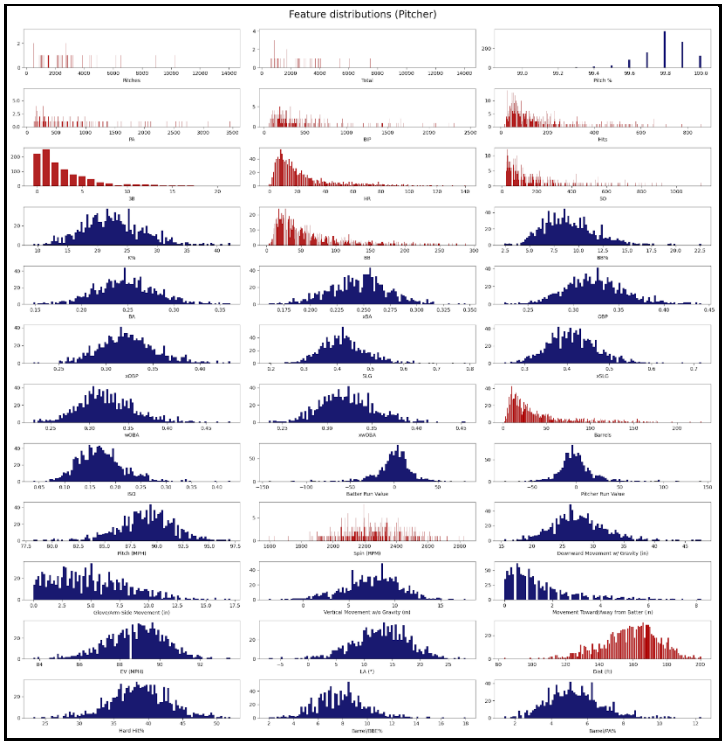
0	count	mean	std	min	5%	25%	50%	75%	90%	95%	99%	max
Pitches	804.00	3,553.68	2,687.48	1,016.00	1,110.15	1,649.75	2,579.50	4,286.50	7,662.40	9,988.90	12,613.69	14,081.00
Total	804.00	3,560.93	2,693.29	1,016.00	1,112.15	1,653.00	2,584.00	4,298.00	7,685.70	10,008.30	12,658.97	14,089.00
Pitch %	804.00	99.77	0.40	96.10	99.10	99.70	99.90	100.00	100.00	100.00	100.00	100.00
PA	804.00	957.67	755.98	109.00	147.30	349.00	690.00	1,485.25	2,174.70	2,472.25	2,831.73	3,180.00
BIP	804.00	648.80	533.61	19.00	91.30	218.75	453.50	1,018.25	1,513.10	1,746.60	2,027.94	2,337.00
Hits	804.00	213.41	182.69	6.00	28.00	68.75	146.00	323.00	503.70	598.40	705.79	839.00
3B	804.00	3.74	4.51	0.00	0.00	1.00	2.00	5.00	9.00	12.85	21.00	31.00
HR	804.00	32.14	34.30	0.00	2.00	7.75	20.00	44.00	87.00	106.85	140.00	192.00
SO	804.00	217.19	162.93	19.00	40.00	83.00	168.50	323.25	460.00	526.65	686.34	837.00
K%	804.00	24.63	6.88	5.10	14.12	20.30	24.40	28.70	32.77	35.67	39.89	77.10
BB	804.00	80.19	75.46	2.00	11.00	24.00	53.00	114.25	189.00	241.55	323.79	506.00
BB%	804.00	8.09	2.70	1.40	4.02	6.10	7.90	9.70	11.60	12.90	15.00	18.00
BA	804.00	0.24	0.03	0.06	0.17	0.22	0.24	0.26	0.27	0.28	0.30	0.33
xBA	804.00	0.23	0.03	0.05	0.19	0.22	0.24	0.25	0.27	0.28	0.29	0.32
OBP	804.00	0.30	0.04	0.11	0.25	0.29	0.31	0.33	0.34	0.36	0.39	0.43
xOBP	804.00	0.31	0.03	0.10	0.25	0.29	0.31	0.33	0.34	0.35	0.39	0.42
SLG	804.00	0.39	0.07	0.08	0.28	0.35	0.39	0.44	0.47	0.50	0.54	0.60
xSLG	804.00	0.39	0.07	0.06	0.29	0.34	0.38	0.43	0.47	0.48	0.55	0.64
wOBA	804.00	0.30	0.04	0.09	0.24	0.28	0.30	0.33	0.35	0.36	0.39	0.42
xwOBA	804.00	0.30	0.04	0.08	0.25	0.28	0.30	0.33	0.35	0.36	0.39	0.43
Barrels	804.00	51.12	55.43	0.00	3.00	11.00	30.00	72.00	134.40	168.00	239.97	297.00
ISO	804.00	0.16	0.05	0.02	0.07	0.12	0.15	0.19	0.22	0.24	0.28	0.33
Batter Run Value	804.00	3.07	34.28	-81.90	-31.85	-15.00	-5.00	9.40	42.10	71.74	139.81	243.50
Pitcher Run Value	804.00	-3.07	34.28	-243.50	-71.74	-9.40	5.00	15.00	24.07	31.85	48.05	81.90
Pitch (MPH)	804.00	88.76	0.46	87.30	88.00	88.50	88.80	89.10	89.30	89.50	89.80	90.10
Spin (RPM)	804.00	2,247.03	27.56	2,148.00	2,203.00	2,229.00	2,248.00	2,264.00	2,279.00	2,288.00	2,314.85	2,405.00
Downward Movement w/ Gravity (in)	804.00	27.87	0.99	24.70	26.30	27.20	27.90	28.50	29.10	29.50	30.49	31.60
Glove/Arm-Side Movement (in)	804.00	4.19	0.78	0.80	3.00	3.60	4.20	4.70	5.17	5.60	5.90	6.80
Vertical Movement w/o Gravity (in)	804.00	7.55	0.73	4.90	6.40	7.10	7.60	8.00	8.50	8.70	9.20	9.90
Movement Toward/Away from Batter (in)	804.00	2.01	1.64	0.00	0.10	0.60	1.45	3.30	4.50	5.00	5.80	6.50
EV (MPH)	804.00	88.31	2.11	81.40	84.80	86.90	88.40	89.72	90.90	91.60	92.70	96.00
LA (°)	804.00	12.60	4.70	-20.90	4.90	9.67	12.80	15.90	18.40	19.50	22.29	26.00
Dist (ft)	804.00	156.35	11.00	123.00	137.15	149.00	157.00	164.00	170.00	173.00	180.00	190.00
Hard Hit%	804.00	37.14	7.39	11.10	24.00	32.40	37.70	42.52	45.90	47.87	51.50	59.20
Barrel/BBE%	804.00	7.29	3.82	0.00	1.70	4.50	7.00	9.80	12.30	14.20	17.20	22.10
Barrel/PA%	804.00	4.69	2.28	0.00	1.10	3.00	4.60	6.30	7.80	8.60	10.00	12.80

รูปที่ 3-10 ข้อมูลสถิติ Batter

จากข้อมูลสถิติของ Pitcher และ Batter จะเห็นได้ว่า Features Pitches, Total และ Spin (RPM) มีค่าที่แตกต่างจาก Features อื่น ๆ อย่างมาก นั่นหมายความว่าทั้ง 2 ชุดข้อมูลนี้มีการกระจายตัวแบบไม่ปกติ และยังพบว่า LA (°) มี Missing Value สำหรับ Pitcher รวมไปถึงค่าที่เป็น 0 ใน Pitcher และ Batter อีกด้วย

3.2.3. Distributions การกระจายตัวของข้อมูล

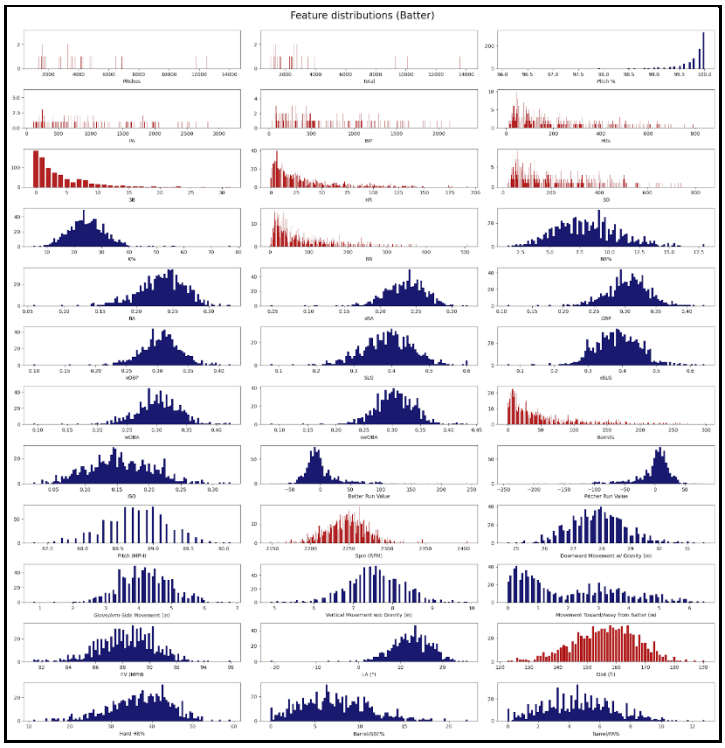
ตรวจสอบการกระจายตัวของข้อมูลด้วยการ Plot กราฟในรูปแบบ Histogram (สีน้ำเงิน) สำหรับข้อมูลประเภท Float และ Plot กราฟ ในรูปแบบของ Bar Chart (สีแดง) สำหรับข้อมูลประเภท Integer ดังกราฟด้านล่าง



รูปที่ 3-11 กราฟการกระจายตัวของข้อมูล Pitcher

จากกราฟการกระจายตัวข้อมูลของ Pitcher จะแสดงให้เห็นว่ามีการแจกแจงไม่ปกติและส่วนใหญ่มีลักษณะเบ้ขวา (Right-Skewed Distribution) รวมถึงมีความแปรปรวนในการกระจายข้อมูล ซึ่งสามารถเกิดได้จากหลายสาเหตุ และยังมีค่า Outliers ในบางจุดของข้อมูล ยกตัวอย่างเช่น ผู้เล่นเบสบอลบางคนสามารถตีโฮมรันได้มากกว่า 100 ในแต่ละฤดูกาล ซึ่งเป็นเหตุการณ์ที่เกิดขึ้นแบบพิเศษ

ในส่วนของ Batter เป็นดังรูปภาพด้านล่างนี้

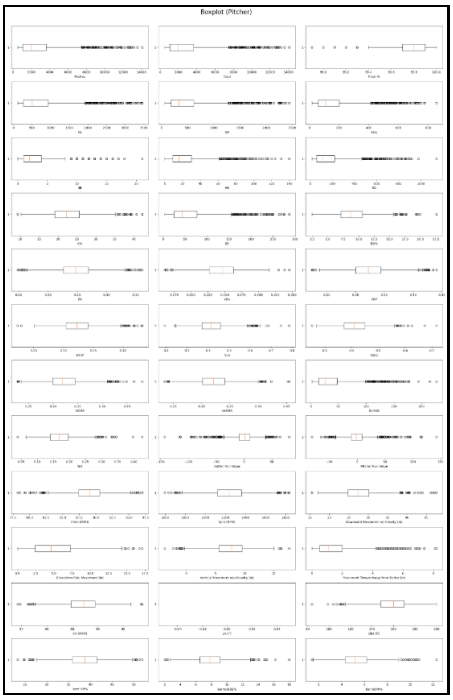


รูปที่ 3-12 กราฟการกระจายตัวของข้อมูล Batter

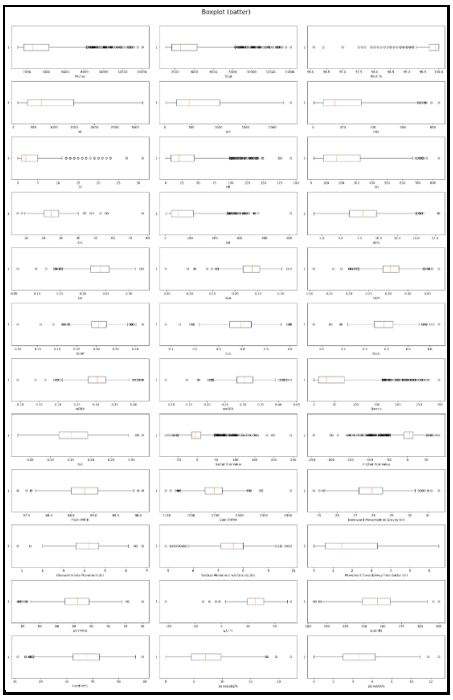
กราฟการกระจายตัวของ Batter จะเห็นได้ว่าข้อมูลส่วนใหญ่มีการกระจายตัวแบบทั้งเบ้ขวา (Right-Skewed Distribution) และเบ้ซ้าย (Left-Skewed Distribution) ยกตัวอย่างเช่น ผู้เล่นเบสบอลที่มีค่าเฉลี่ยการตี (BA) สูงนั้นหายากกว่าผู้เล่นเบสบอลที่มีค่าเฉลี่ยการตีต่ำ นอกจากนี้ข้อมูลยังมีค่า Outliers ที่ต่างจากค่าทั่วไปมาก เช่น HR (Home runs) และ BB (Bases on Balls) สาเหตุอาจมาจากผู้เล่นตีโฮมรันหลายครั้งหรือทำคะแนนได้จำนวนมาก เป็นต้น

3.2.4. Outliers

จะแสดงข้อมูลอยู่ในรูปของกราฟ Box plot เพื่อดูการกระจายตัวของข้อมูลและตรวจสอบ Outliers ของแต่ละ Features ในชุดข้อมูลของ Pitcher กับ Batter ซึ่งจะช่วยให้สามารถวิเคราะห์ข้อมูลได้ง่ายขึ้น ดังตามรูปภาพด้านล่างนี้



รูปที่ 3-13 กราฟ Box Plot ของข้อมูล Pitcher

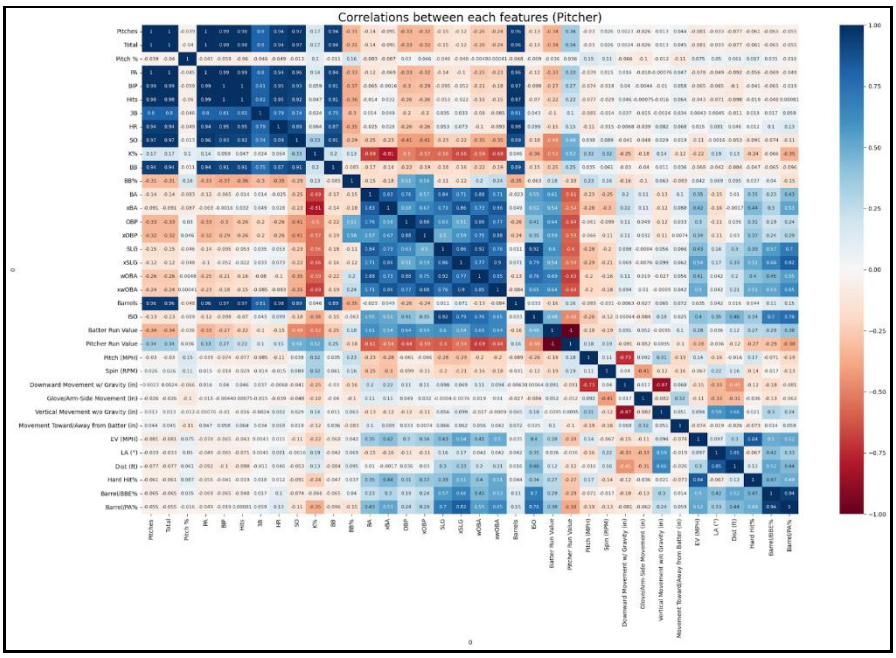


รูปที่ 3-14 กราฟ Box Plot ของข้อมูล Batter

พบว่า ทั้งชุดข้อมูล Pitcher และ Batter มีจำนวน Outliers เป็นจำนวนมากในแต่ละ Feature จะสังเกตได้จาก Data Points ที่อยู่นอก Upper Whisker และ Lower Whisker ของ Box Plot

3.2.5. Correlations

ทำการแสดงความสัมพันธ์ (Correlations) แต่ละ Features ของข้อมูล Pitcher กับ Batter และแสดงผลความสัมพันธ์ในรูปของ Heatmap ซึ่งสีที่เข้มบ่งบอกถึงความสัมพันธ์ที่สูงและสีที่อ่อนบ่งบอกถึงความสัมพันธ์ที่ต่ำ โดยเทียบกับเป้าหมาย (Pitches) เนื่องจาก ถ้าจำนวนครั้งที่ขว้างหรือตีได้เยอะ ก็หมายความว่า ผู้เล่นมีสามารถในการทำคะแนนได้เยอะ



รูปที่ 3-15 กราฟ Correlation ของข้อมูล Pitcher

จากกราฟ Heatmap สามารถบอกความสัมพันธ์เบื้องต้นระหว่าง Features กับ Target (Pitches) ได้ของชุดข้อมูล Pitcher ดังนี้

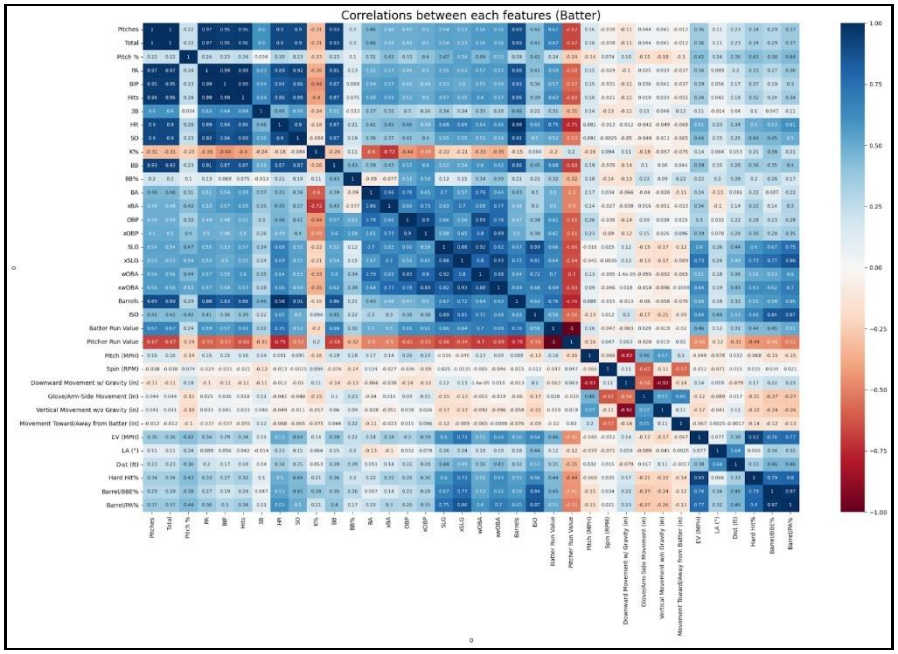
- BIP, Hits, 3B, HR, SO, BB และ Barrels มีความสัมพันธ์เชิงบวกที่ค่อนข้างสูงมากต่อเป้าหมาย (Strong Positive Relationship) หมายความว่า คุณสมบัติเหล่านี้มีแนวโน้มที่จะเพิ่มขึ้นพร้อมกับจำนวน Pitches ที่ขว้างสูง
- BB%, OBP, xOBP และ Batter Run Value มีความสัมพันธ์เชิงลบที่ค่อนข้างสูงกว่าฟีเจอร์อื่น ๆ (Moderate Negative Relationship) หมายความว่า คุณสมบัติเหล่านี้มีแนวโน้มที่จะลดลงพร้อมกับจำนวน Pitches ที่ขว้างในระดับปานกลาง
- Pitch%, K%, BA, xBA, SLG, xSLG, wOBA, xwOBA, ISO, Pitch(MPH), Spin(RPM), Downward Movement w/ Gravity (in), Glove/Arm-Side Movement (in), Vertical Movement w/o Gravity (in), Movement Toward/Away from Batter (in), EV (MPH), LA (°), Dist (ft),

Hard Hit%, Barrel/BBE%, Barrel/PA% มีความสัมพันธ์ Weak Correlations ต่อเป้าหมาย หมายความว่า คุณสมบัติเหล่านี้มีความสัมพันธ์เชิงเส้นน้อยมากกับจำนวน Pitches ที่ขว้าง

- Total และ PA มีความสัมพันธ์ Perfectly Positive Correlations ต่อเป้าหมาย หมายความว่า คุณสมบัติเหล่านี้มีความสัมพันธ์เชิงเส้นสมบูรณ์กับจำนวน Pitches ที่ขว้าง

โดยสรุปแล้ว ความสัมพันธ์เชิงบวกที่ค่อนข้างสูงมากระหว่าง BIP, Hits, 3B, HR, SO, BB และ Barrels กับเป้าหมาย (Pitches) บ่งชี้ได้เบื้องต้นว่าฟีเจอร์เหล่านี้เป็นตัวทำนายประสิทธิภาพการขว้างลูกของนักขว้างที่ดี

ความสัมพันธ์ระหว่าง Features ของข้อมูล Batter ดังรูปภาพด้านล่างนี้



รูปที่ 3-16 กราฟ Correlation ของข้อมูล Batter

ซึ่งสามารถแสดงให้เห็นความสัมพันธ์ระหว่าง Features กับ Target (Pitches) ของชุดข้อมูล Batter ได้ดังนี้

- PA, BIP, Hits, HR, SO, BB และ Barrels มีความสัมพันธ์เชิงบวกที่ค่อนข้างสูงมากต่อเป้าหมาย (Strong Positive Relationship)
- K% และ Pitcher Run Value มีความสัมพันธ์เชิงลบที่ค่อนข้างสูงกว่าฟีเจอร์อื่น ๆ (Moderate Negative - Relationship) หมายความว่า ฟีเจอร์เหล่านี้มีแนวโน้มที่จะลดลงพร้อมกับจำนวน Pitches ที่ตีในระดับปานกลาง

- Pitch%, BB%, Pitch(MPH), Spin(RPM), Downward Movement w/Gravity(in), Glove/Arm-Side Movement(in), Vertical Movement w/o Gravity(in), Movement Toward/Away from Batter(in), LA (°), Dist(ft) และ Barrel/BBE% มีความสัมพันธ์ Weak Correlations ต่อเป้าหมาย หมายความว่า คุณสมบัติเหล่านี้มีความสัมพันธ์เชิงเส้นน้อยมากกับจำนวน Pitches ที่ตี
- Total มีความสัมพันธ์ Perfectly Positive Correlations ต่อเป้าหมาย หมายความว่า คุณสมบัติเหล่านี้มีความสัมพันธ์เชิงเส้นสมบูรณ์กับจำนวน Pitches ที่ตี

3.3. การเตรียมข้อมูลเบื้องต้น (Data Preprocessing / Data Cleaning)

เป็นขั้นตอนสำคัญในการเตรียมข้อมูลเพื่อให้ข้อมูลเหมาะสมและมีคุณภาพก่อนที่จะนำไปวิเคราะห์และสร้างโมเดล

3.3.1. Handle Missing Value

จะทำการจัดการค่า 0 ในข้อมูลชุด Pitcher และ Batter โดยจะใช้คำสั่ง pitcher.eq(0).sum() ในการนับจำนวนครั้งที่ค่าเป็น 0 ในแต่ละคอลัมน์ จากนั้นจะทำการแทนค่า 0 ด้วยค่า NaN ซึ่งเป็นค่าที่ใช้แทนค่าที่หายไปหรือไม่มีความหมาย เพื่อไม่ให้ค่าที่เป็น 0 มีผลกระทบต่อโมเดลและช่วยปรับปรุงประสิทธิภาพของโมเดล โดยใช้คำสั่ง pitcher.replace(0, np.nan, inplace=True) ดังโค้ดด้านล่างนี้

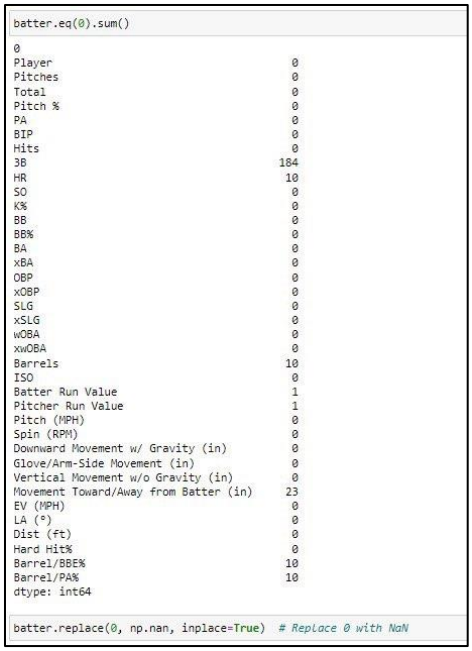
3.1 Remove Missing Values

```
pitcher.eq(0).sum()
```

```
0
Player                                0
Pitches                               0
Total                                 0
Pitch %                               0
PA                                    0
BIP                                   0
Hits                                  0
3B                                  221
HR                                    3
SO                                    0
K%                                    0
BB                                    0
BB%                                   0
BA                                    0
xBA                                   0
OBP                                   0
xOBP                                  0
SLG                                   0
xSLG                                  0
wOBA                                  0
xwOBA                                 0
Barrels                               0
ISO                                   0
Batter Run Value                      5
Pitcher Run Value                     5
Pitch (MPH)                           0
Spin (RPM)                            0
Downward Movement w/ Gravity (in)     0
Glove/Arm-Side Movement (in)          9
Vertical Movement w/o Gravity (in)    1
Movement Toward/Away from Batter (in)    35
EV (MPH)                              0
LA (°)                                0
Dist (ft)                             0
Hard Hit%                             0
Barrel/BBE%                           0
Barrel/PA%                            0
dtype: int64
```

```
pitcher.replace(0, np.nan, inplace=True)  # Replace 0 with NaN
```

รูปที่ 3-17 แทนค่า 0 ด้วย NaN ของ Pitcher



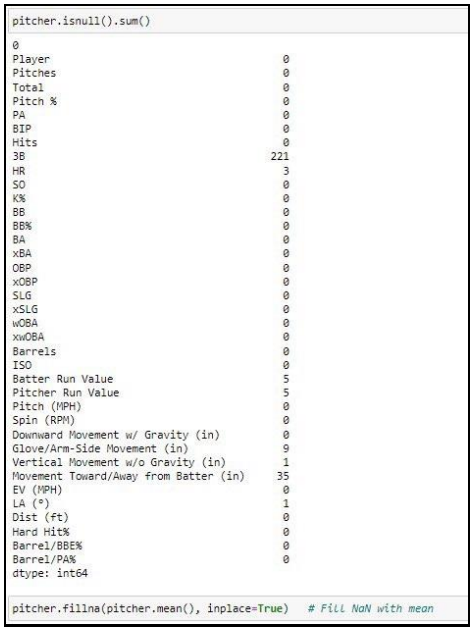
```
batter.eq(0).sum()
```

```
0
Player                                0
Pitches                               0
Total                                 0
Pitch %                               0
PA                                    0
BIP                                   0
Hits                                  0
3B                                  184
HR                                   10
SO                                    0
K%                                    0
BB                                    0
BB%                                   0
BA                                    0
xBA                                   0
OBP                                   0
xOBP                                  0
SLG                                   0
xSLG                                  0
wOBA                                  0
xwOBA                                 0
Barrels                              10
ISO                                   0
Batter Run Value                      1
Pitcher Run Value                     1
Pitch (MPH)                           0
Spin (RPM)                            0
Downward Movement w/ Gravity (in)     0
Glove/Arm-Side Movement (in)          0
Vertical Movement w/o Gravity (in)    0
Movement Toward/Away from Batter (in)    23
EV (MPH)                              0
LA (°)                                0
Dist (ft)                             0
Hard Hit%                             0
Barrel/BBE%                          10
Barrel/PA%                           10
dtype: int64
```

```
batter.replace(0, np.nan, inplace=True)  # Replace 0 with NaN
```

รูปที่ 3-18 แทนค่า 0 ด้วย NaN ของ Batter

จากนั้นจะใช้คำสั่ง pitcher.isnull().sum() เพื่อนับจำนวนค่า NaN หลังจากที่ค่า 0 ถูกแทนด้วยค่า NaN แล้วทำการแทนที่ค่า NaN ด้วยค่าเฉลี่ยของแต่ละคอลัมน์ ตัวอย่างเช่น คอลัมน์ A มีค่าเป็น [1, 2, NaN, 4, 5] ค่า NaN ในตำแหน่งที่ 3 จะถูกแทนที่ด้วยค่าเฉลี่ยของ [1, 2, 4, 5] ซึ่งคือ (1+2+4+5)/4 = 3.0 ดังนั้นคอลัมน์ A จะกลายเป็น [1, 2, 3.0, 4, 5] โดยจะใช้คำสั่ง pitcher.fillna(pitcher.mean(), inplace = True) ดังโค้ดด้านล่าง

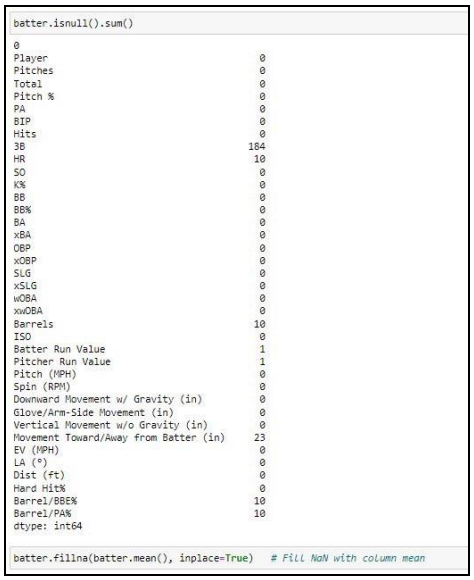


```
pitcher.isnull().sum()
```

```
0
Player                                0
Pitches                               0
Total                                 0
Pitch %                               0
PA                                    0
BIP                                   0
Hits                                  0
3B                                  221
HR                                    3
SO                                    0
K%                                    0
BB                                    0
BB%                                   0
BA                                    0
xBA                                   0
OBP                                   0
xOBP                                  0
SLG                                   0
xSLG                                  0
wOBA                                  0
xwOBA                                 0
Barrels                               0
ISO                                   0
Batter Run Value                      5
Pitcher Run Value                     5
Pitch (MPH)                           0
Spin (RPM)                            0
Downward Movement w/ Gravity (in)     0
Glove/Arm-Side Movement (in)          9
Vertical Movement w/o Gravity (in)    1
Movement Toward/Away from Batter (in)    35
EV (MPH)                              0
LA (°)                                1
Dist (ft)                             0
Hard Hit%                             0
Barrel/BBE%                           0
Barrel/PA%                            0
dtype: int64
```

```
pitcher.fillna(pitcher.mean(), inplace=True)   # Fill NaN with mean
```

รูปที่ 3-19 แทนค่า NaN ด้วยค่าเฉลี่ย ของ Pitcher



```
batter.isnull().sum()
```

```
0
Player                                0
Pitches                               0
Total                                 0
Pitch %                               0
PA                                    0
BIP                                   0
Hits                                  0
3B                                  184
HR                                   10
SO                                    0
K%                                    0
BB                                    0
BB%                                   0
BA                                    0
xBA                                   0
OBP                                   0
xOBP                                  0
SLG                                   0
xSLG                                  0
wOBA                                  0
xwOBA                                 0
Barrels                              10
ISO                                   0
Batter Run Value                      1
Pitcher Run Value                     1
Pitch (MPH)                           0
Spin (RPM)                            0
Downward Movement w/ Gravity (in)     0
Glove/Arm-Side Movement (in)          0
Vertical Movement w/o Gravity (in)    0
Movement Toward/Away from Batter (in)    23
EV (MPH)                              0
LA (°)                                0
Dist (ft)                             0
Hard Hit%                             0
Barrel/BBE%                          10
Barrel/PA%                           10
dtype: int64
```

```
batter.fillna(batter.mean(), inplace=True)   # Fill NaN with column mean
```

รูปที่ 3-20 แทนค่า NaN ด้วยค่าเฉลี่ย ของ Batter

3.4. การสร้างแบบจำลอง (Model Building)

ทำการลบ Features ที่ไม่จำเป็นหรือคล้ายกันออกของข้อมูล Pitcher และทำการเก็บข้อมูลในรูปแบบ DataFrame ที่มีชื่อว่า train_X เพื่อใช้สำหรับการ Train Model

ในส่วนของข้อมูล Batter ก็ทำการลบ Features ที่ไม่จำเป็นแล้วทำการเก็บข้อมูลในรูปแบบ DataFrame ที่มีชื่อว่า train_X_batter

```
train_X = pitcher.drop(['Player', 'Total', 'Pitch %', 'BB%', 'xBA', 'xOBP', 'xSLG', 'xwOBA'], axis=1).copy()
```

```
train_X=scale(train_X)
```

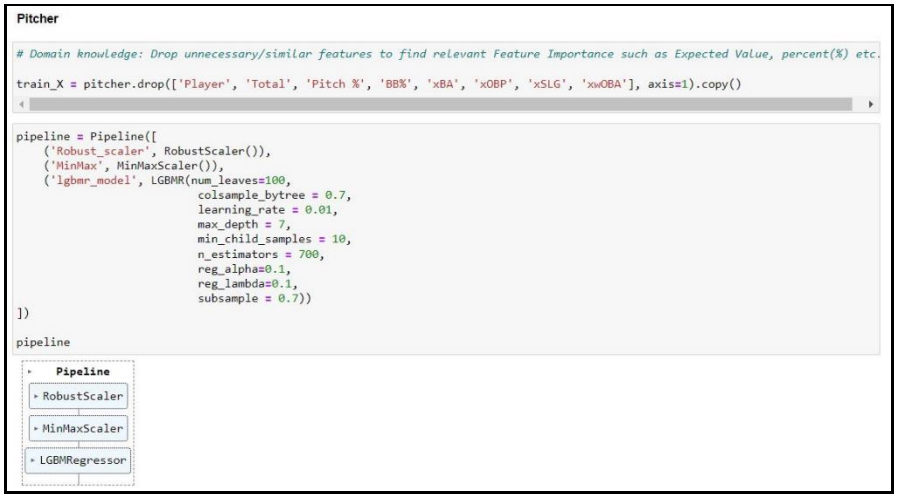
รูปที่ 3-21 ตัวอย่างการ Drop Feature ของ Pitcher

3.4.1. Pipeline

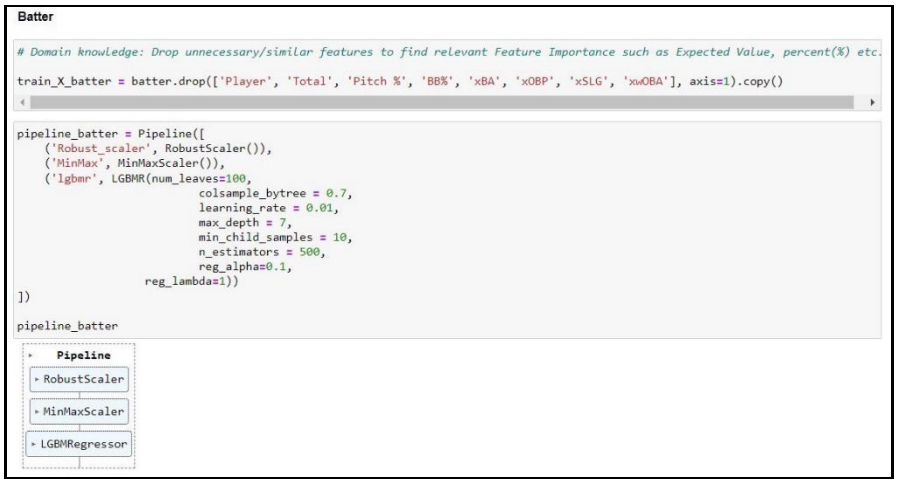
Pipeline ใช้สร้างกระบวนการที่ทำงานต่อเนื่องตั้งแต่การเตรียมข้อมูล (Preprocessing) จนถึงการสร้างและ Training Model เป็นโครงสร้างที่มีประสิทธิภาพในการจัดการขั้นตอนทั้งหมดในการพัฒนาและประเมินโมเดล

โดยจะใช้ RobustScaler เพื่อปรับสเกลข้อมูลที่ทนทานต่อ Outliers สามารถทำให้ข้อมูลปรับสเกลได้อย่างมีประสิทธิภาพโดยไม่ต้องได้รับผลกระทบจากค่า Outliers และใช้ MinMaxScaler เพื่อปรับค่าของข้อมูลให้อยู่ในช่วง [0, 1] โดยเลือกทำ RobustScaler ก่อน เพื่อให้ MinMaxScaler ที่ Sensitive ต่อ Outliers มีผลน้อยลง

จากนั้นจะใช้ LightGBMRegressor เป็นอัลกอริทึมที่ใช้ในการทำนายแบบ Regression เพื่อทำการสร้าง Model ทั้งของ Pitcher และ Batter ซึ่ง LightGBM เป็นโมเดลการเรียนรู้แบบ Gradient Boosting ที่ใช้ Tree - Based Learning (เทคนิคการเรียนรู้ที่ใช้ Decision Trees) มีจุดเด่นที่ความเร็วในการ Train และประสิทธิภาพสูง



รูปที่ 3-22 Pipeline ของ Pitcher



รูปที่ 3-23 Pipeline ของ Batter

3.4.2. Training Model

ทำการแบ่งชุดข้อมูลสำหรับ Pitcher และ Batter ออกเป็นข้อมูล Train set และ Test set โดยแบ่งให้มีอัตราส่วน 80% และ 20% ตามลำดับ นำข้อมูลที่ใช้ในการ Train Model ทำ Cross-Validation ในการแบ่งข้อมูล ด้วยเทคนิค K-Fold (กำหนด K=10) รวมไปถึงการทำ Feature Selection เลือกใช้เทคนิค Recursive Feature Selection (RFE) เพื่อช่วยลดจำนวน Features ที่ไม่มีความสำคัญ หรือมีความสำคัญน้อย โดยกำหนดเลือกให้เหลือเป็นจำนวน 10 Features ซึ่งการทำ Cross-Validation และ Feature Selection ช่วยลดปัญหา Overfitting และปรับปรุงประสิทธิภาพของโมเดลให้มีประสิทธิภาพมากยิ่งขึ้นได้

จากนั้นนำข้อมูลที่ผ่านกระบวนการใน Pipeline ที่มี LightGBMRegressor อัลกอริทึมในการสร้างโมเดล นำไปผ่านการ Train และ Evaluate model ด้วยค่า Mean Absolute Error (MAE) สำหรับคำนวณหาค่าเฉลี่ยของแต่ละ Fold เพื่อประเมินประสิทธิภาพโดยรวมของโมเดล

Pitcher

```
X = train_X.drop(['Pitches'], axis = 1)
y=pitcher["Pitches"]
```

```
n_selected_features = 10

# Initialize RFE
rfe = RFE(estimator=pipeline.named_steps['lgbmr_model'], n_features_to_select=n_selected_features)


# Split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize KFold for cross-validation within the training set
kf = KFold(n_splits=10, shuffle=True, random_state=42)

# Initialize a list to store mean absolute errors for each fold
mae_scores = []

# Iterate over the folds
for fold, (train_idx, val_idx) in enumerate(kf.split(X_train, y_train), 1):
    X_train_fold, X_val_fold = X_train.iloc[train_idx], X_train.iloc[val_idx]
    y_train_fold, y_val_fold = y_train.iloc[train_idx], y_train.iloc[val_idx]

    # Fit RFE on training data within the fold and transform both training and validation data
    X_train_rfe = rfe.fit_transform(X_train_fold, y_train_fold)
    X_val_rfe = rfe.transform(X_val_fold)

    # Train the pipeline on the training data with selected features
    pipeline.fit(X_train_rfe, y_train_fold)

    # Make predictions on the validation set
    y_pred = pipeline.predict(X_val_rfe)


    # Evaluate the performance with mean absolute error
    mae = mean_absolute_error(y_val_fold, y_pred)
    print(f'Fold {fold}: Mean Absolute Error: {mae}')
    mae_scores.append(mae)

# After cross-validation, fit RFE on the entire training set and transform train and test sets
X_train_rfe = rfe.fit_transform(X_train, y_train)
X_test_rfe = rfe.transform(X_test)

# Train the pipeline on the entire training set with selected features
pipeline.fit(X_train_rfe, y_train)

# Evaluate final performance on the test set
y_pred_test = pipeline.predict(X_test_rfe)
```

รูปที่ 3-24 การ Training Model ของ Pitcher

Batter

```
X_batter = train_X_batter.drop(['Pitches'], axis = 1)
y_batter=batter["Pitches"]
```

```
n_selected_features_batter = 10

# Initialize RFE
rfe_batter = RFE(estimator=pipeline_batter.named_steps['lgbmr'], n_features_to_select=n_selected_features_batter)


# Split data into train and test sets
X_train_batter, X_test_batter, y_train_batter, y_test_batter = train_test_split(X_batter, y_batter, test_size=0.2,
                                                                                random_state=42)

# Initialize KFold for cross-validation within the training set
kf_batter = KFold(n_splits=10, shuffle=True, random_state=42)

# Initialize a list to store mean absolute errors for each fold
mae_scores_batter = []

# Iterate over the folds
for fold, (train_idx_batter, val_idx_batter) in enumerate(kf_batter.split(X_train_batter, y_train_batter), 1):
    X_train_fold_batter, X_val_fold_batter = X_train_batter.iloc[train_idx_batter], X_train_batter.iloc[val_idx_batter]
    y_train_fold_batter, y_val_fold_batter = y_train_batter.iloc[train_idx_batter], y_train_batter.iloc[val_idx_batter]

    # Fit RFE on training data within the fold and transform both training and validation data
    X_train_rfe_batter = rfe_batter.fit_transform(X_train_fold_batter, y_train_fold_batter)
    X_val_rfe_batter = rfe_batter.transform(X_val_fold_batter)

    # Train the pipeline on the training data with selected features
    pipeline_batter.fit(X_train_rfe_batter, y_train_fold_batter)

    # Make predictions on the validation set
    y_pred_batter = pipeline_batter.predict(X_val_rfe_batter)


    # Evaluate the performance with mean absolute error
    mae_batter = mean_absolute_error(y_val_fold_batter, y_pred_batter)
    print(f'Fold {fold}: Mean Absolute Error: {mae_batter}')
    mae_scores_batter.append(mae_batter)

# After cross-validation, fit RFE on the entire training set and transform train and test sets
X_train_rfe_batter = rfe_batter.fit_transform(X_train_batter, y_train_batter)
X_test_rfe_batter = rfe_batter.transform(X_test_batter)

# Train the pipeline on the entire training set with selected features
pipeline_batter.fit(X_train_rfe_batter, y_train_batter)

# Evaluate final performance on the test set
y_pred_test_batter = pipeline_batter.predict(X_test_rfe_batter)
```

รูปที่ 3-25 การ Training Model ของ Batter

ซึ่งจะได้ค่า Average Mean Absolute Error ที่ได้จากการประเมินโมเดลทั้งหมดบนทุก Fold ของชุดข้อมูล Test ของ Pitcher แล้วนำมาเฉลี่ยกันอยู่ที่ประมาณ 66.29 ถือเป็นค่า MAE ที่ค่อนข้างต่ำ และแสดงถึงความแม่นยำของโมเดลที่ดีในการทำนายข้อมูลในระดับนึง

Pitcher

```
final_mae = mean_absolute_error(y_test, y_pred_test)
print(f'MAE of Pitcher: {final_mae}')
```

```
MAE of Pitcher: 66.2935731661671
```

รูปที่ 3-26 ค่า Average Mean Absolute Error ของ Pitcher

สำหรับชุดข้อมูล Batter ค่า Average Mean Absolute Error จะอยู่ที่ประมาณ 121.28 ซึ่งถือว่าเป็นค่า MAE ที่สูง เมื่อเทียบกับค่า MAE ของ Pitcher และจำนวนข้อมูลที่มี ซึ่งคณะผู้จัดทำมีความคิดเห็นว่า อาจจะเกิดจากการที่มีข้อมูลไม่เพียงพอ ทำให้ Train Model ไม่มีความหลากหลายในการเรียนรู้หรือเรียนรู้ได้น้อย

Batter

```
final_mae_batter = mean_absolute_error(y_test_batter, y_pred_test_batter)
print(f'MAE of Batter: {final_mae_batter}')
```

```
MAE of Batter: 121.27667986615411
```

รูปที่ 3-27 ค่า Average Mean Absolute Error ของ Batter

3.4.3. Feature Importance

แสดงผลลัพธ์ที่ได้จากโมเดล เพื่อบอกถึงความสำคัญของแต่ละ Feature ที่มีผลต่อการทำนาย หรือก็คือ Features ที่มีผลต่อเป้าหมาย (Pitches) โดยจัดเรียงตามลำดับความสำคัญและแสดงผลในรูปแบบของกราฟ ตามโค้ดด้านล่างนี้

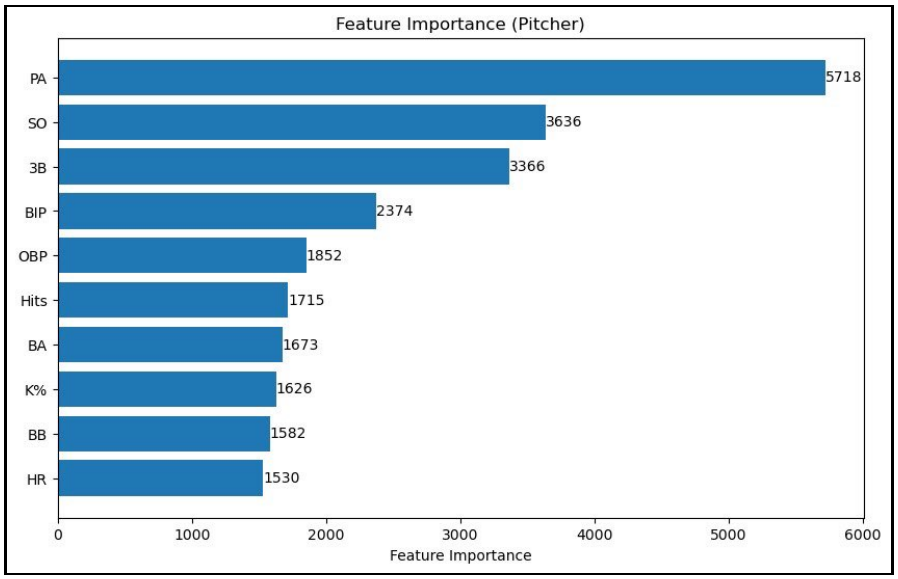
```
feature_importances = pipeline.named_steps['lgbmr_model'].feature_importances_

sorted_idx = feature_importances.argsort()

plt.figure(figsize=(10, 6))
plt.barh(range(len(sorted_idx)), feature_importances[sorted_idx], align="center")

for i, v in enumerate(feature_importances[sorted_idx]):
    plt.text(v + 0.01, i, str(round(v, 2)), ha='left', va='center')

plt.yticks(range(len(sorted_idx)), X.columns[sorted_idx])
plt.xlabel("Feature Importance")
plt.title("Feature Importance (Pitcher)")

plt.show()
```

รูปที่ 3-28 Feature Importance ของ Pitcher

```
feature_importances_batter = pipeline_batter.named_steps['lgbmr'].feature_importances_

sorted_idx_batter = feature_importances_batter.argsort()

plt.figure(figsize=(10, 6))
plt.barh(range(len(sorted_idx_batter)), feature_importances_batter[sorted_idx_batter], align="center")

for i, v in enumerate(feature_importances_batter[sorted_idx_batter]):
    plt.text(v + 0.01, i, str(round(v, 2)), ha='left', va='center')

plt.yticks(range(len(sorted_idx_batter)), X_batter.columns[sorted_idx_batter])
plt.xlabel("Feature Importance")
plt.title("Feature Importance (Batter)")

plt.show()
```

รูปที่ 3-29 Feature Importance ของ Batter

โดย Feature Importance ของชุดข้อมูล Pitcher จะได้ออกมาดังนี้



รูปที่ 3-30 กราฟ Feature Importance ของ Pitcher

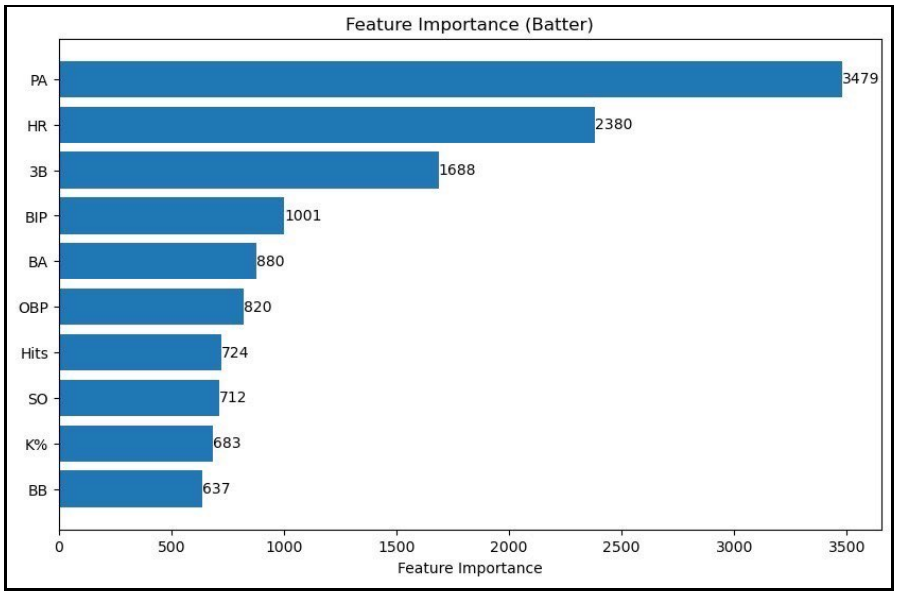
จากกราฟจะสรุปได้ว่า 10 อันดับ Features ที่ส่งผลต่อเป้าหมาย (Pitches) ของชุดข้อมูล Pitcher มากที่สุดจะเป็น PA, SO, 3B, BIP, OBP, Hits, BA, K%, BB และ HR ตามลำดับ

จะเห็นได้ว่าบาง Features มีความสัมพันธ์ Correlation ต่อ Pitches แบบ Weak Correlation เช่น BA, K% เป็นต้น ซึ่งหมายความว่า ฟีเจอร์เหล่านี้มีความสัมพันธ์เชิงเส้นน้อยมากกับเป้าหมาย

นั่นทำให้สรุปได้ว่า ความสัมพันธ์ไม่ได้เป็นตัวกำหนดเสมอไปว่า Features เหล่านี้เป็นปัจจัยสำคัญต่อเป้าหมายหรือไม่ ยกตัวอย่างเช่น ในกรณี BA มีความสัมพันธ์แบบ Weak Correlation กับเป้าหมาย (Pitches) อย่างไรก็ตาม ฟีเจอร์นี้อาจมีความสำคัญเนื่องจากมีความสัมพันธ์เชิงบวกสูงกับคุณสมบัติอื่น ๆ เช่น BIP, Hits, HR และ SO เป็นต้น

ดังนั้น BA หรือ Features อื่น ๆ จึงเป็นปัจจัยสำคัญในการอธิบายเป้าหมาย (Pitches) แม้ว่าจะมีความสัมพันธ์แบบ Weak Correlation ต่อ Target ก็ตาม

สำหรับ Feature Importance ของชุดข้อมูล Batter จะได้ Features ออกมาดังนี้



รูปที่ 3-31 กราฟ Feature Importance ของ Batter

จะได้ว่า Features ที่มีความสำคัญต่อการทำนาย หรือต่อ Target (Pitches) สูงสุด 10 อันดับแรกจะเป็น PA, HR, 3B, BIP, BA, OBP, Hits, SO, K% และ BB ตามลำดับ

3.5. การนำแบบจำลองไปใช้งาน (Model Deployment)

สามารถใช้ประโยชน์จาก Feature Importance ในการหาปัจจัยที่มีผลต่อโอกาสชนะ โดยเทียบกับเป้าหมายของเรา หรือ Pitches

โดยปัจจัยที่มีค่า Feature Importance สูง แสดงว่าปัจจัยนั้นมีความเกี่ยวข้องหรือมีความสัมพันธ์กับเป้าหมาย (Pitches) มาก ปัจจัยเหล่านี้จึงเป็นปัจจัยสำคัญที่ควรพิจารณา ตัวอย่างเช่น จาก Feature Importance ของชุดข้อมูล Pitcher ได้แก่ SO, BB และ BIP ปัจจัยเหล่านี้ล้วนเป็นปัจจัยที่วัดประสิทธิภาพการขว้างลูกของ Pitcher

นอกจากนี้ยังสามารถใช้ Feature Importance ในการเปรียบเทียบปัจจัยต่าง ๆ กัน ตัวอย่างเช่น สามารถเปรียบเทียบค่า Feature Importance ของ BIP (จำนวนครั้งที่ตีลูกได้และสร้างโอกาสในการทำคะแนน) และ Hits (จำนวนครั้งที่ตีได้) ในชุดข้อมูล Batter ผลการเปรียบเทียบแสดงให้เห็นว่า BIP มีความสัมพันธ์กับเป้าหมายมากกว่า Hits นั่นหมายความว่า BIP เป็นปัจจัยที่สำคัญกว่า Hits ในการพัฒนาเพื่อทำให้มีโอกาสชนะมากขึ้น

และยังสามารถนำไปใช้ประโยชน์ในด้านต่าง ๆ ได้ ดังนี้

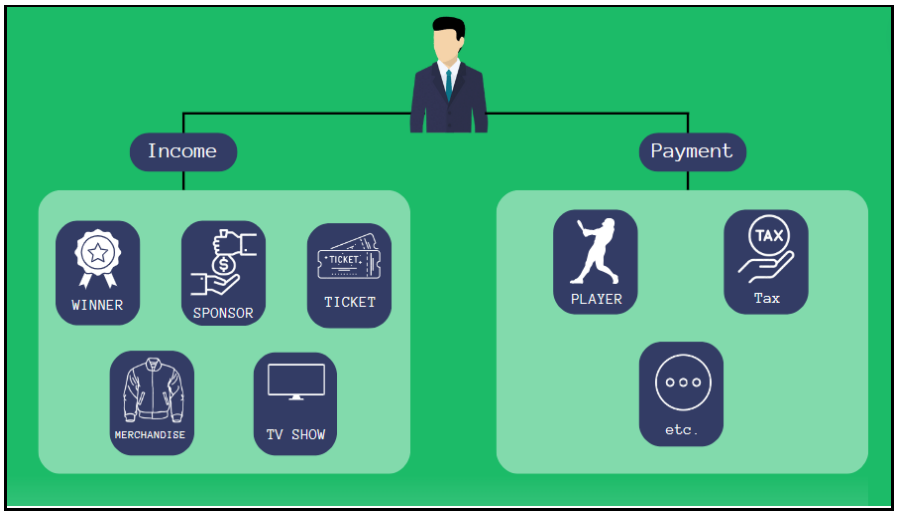
3.5.1. ด้านนักกีฬาและทีม

สามารถใช้เป็นแนวทางในการพัฒนาทักษะและความสามารถของนักกีฬา ปัจจัยที่มีค่า Feature Importance สูง แสดงว่าปัจจัยนั้นมีความสำคัญต่อโอกาสชนะของนักกีฬา นักกีฬาจึงควรให้ความสำคัญในการพัฒนาทักษะและความสามารถในปัจจัยเหล่านี้ เช่น Pitcher ควรพัฒนาทักษะ SO, 3B, BIP เป็นต้น

รวมถึงในด้านการบริหารจัดการทีม สามารถนำไปใช้ประโยชน์ได้ เช่น การกำหนดเป้าหมาย การสร้างแผนการฝึกซ้อม เป็นต้น เพื่อพัฒนาศักยภาพของนักกีฬาและทีมได้อย่างมีประสิทธิภาพ

3.5.2. ด้านธุรกิจและการวิเคราะห์การตลาด

สามารถนำไปช่วยในวางแผนและกำหนดกลยุทธ์ได้ และช่วยในการตัดสินใจได้ดียิ่งขึ้น ยกตัวอย่างเช่น การจัดสรรทรัพยากร หรืองบประมาณ รวมถึงช่วยให้สามารถสร้างกำไรให้แก่องค์กรของผู้บริหารจัดการทีมได้มากขึ้น



รูปที่ 3-32 Income and Payment Baseball

จากรูปจะเห็นได้ว่า รายได้หลักขององค์กรจะมาจาก การที่เราชนะการแข่งขัน (Winner), Sponsor, Ticket, Merchandise และ TV Show ซึ่งการที่เราจะสามารถมี Sponsor, Ticket, Merchandise และ TV Show นั่นแปลว่า เราต้องเป็นทีมที่มีชื่อเสียง โดยการที่เราจะมีชื่อเสียงได้ คือการที่เราต้องชนะการแข่งขัน เพราะฉะนั้นการที่เราหา Feature Importance ที่ส่งผลต่อโอกาสในการชนะจะสามารถตอบโจทย์ทางธุรกิจและสามารถสร้างกำไรให้แก่ผู้บริหารจัดการทีมนั่นเอง

ส่วนในด้านการวิเคราะห์การตลาด สามารถนำข้อมูลที่เราได้มาวิเคราะห์และนำไปพัฒนาแผนการตลาดโดยใช้เครื่องมือทางการตลาดต่าง ๆ ไม่ว่าจะเป็น การตั้งคำถามด้วยหลักการ SMART และ Marketing Analytics

1. การตั้งคำถามด้วยหลักการ SMART

ก่อนการทำ Marketing Analytics จะต้องมีการตั้งคำถาม เพื่อช่วยให้สามารถเข้าใจเป้าหมายและบรรลุได้มีประสิทธิภาพมากขึ้น

โดยเป้าหมาย คือ การหาปัจจัยที่ทำให้มีโอกาสชนะในกีฬาเบสบอลมากขึ้น

Specific (เฉพาะเจาะจง) - ปัจจัยใดบ้างที่ส่งผลต่อโอกาสชนะในเบสบอล

Measurable (วัดผลได้) - คิดว่าจะพัฒนาปัจจัยเหล่านั้นได้อย่างไร

Achievable (ปฏิบัติได้) - สามารถวิเคราะห์ข้อมูลเหล่านั้นได้อย่างมีประสิทธิภาพหรือไม่

Relevant (เกี่ยวข้อง) - ปัจจัยเหล่านี้เกี่ยวข้องโดยตรงกับโอกาสชนะในเบสบอลหรือไม่

Time-bound (กำหนดเวลา) - จะทำอย่างไรให้สามารถพัฒนากลยุทธ์เพื่อพัฒนาปัจจัยต่าง ๆ ได้ ภายในระยะเวลา 1 ปี

2. Marketing Analytics

Descriptive Analytics - การวิเคราะห์ข้อมูลสถิติของทีมเบสบอล เพื่อทราบศักยภาพของทีมและนักกีฬาในปัจจุบัน เช่น อัตราการชนะ อัตราการตีลูก อัตราการขว้างลูก เป็นต้น

Diagnostic Analytics - วิเคราะห์หาสาเหตุว่าทำอย่างไรถึงจะช่วยเพิ่มโอกาสชนะให้กับทีมมากขึ้น

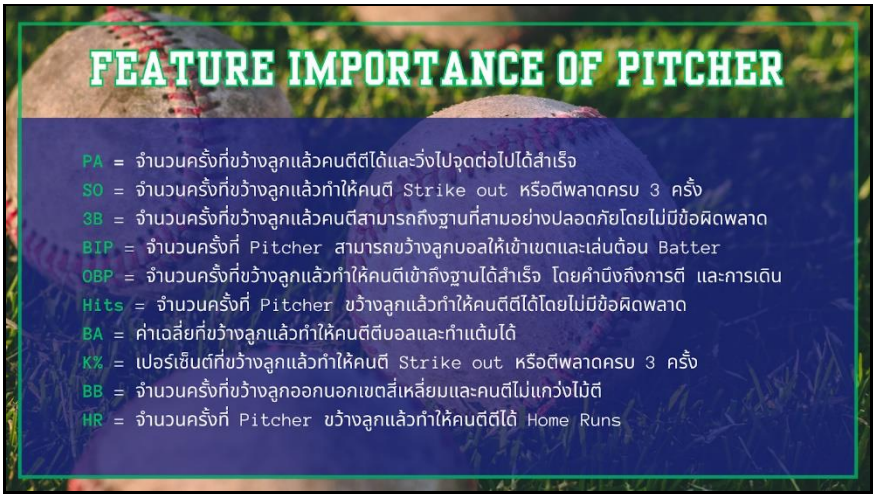
Predictive Analytics - สร้างโมเดลเพื่อหา Feature Importance ที่ส่งผลต่อการชนะของทีม เพื่อช่วยให้ทีมสามารถนำไปใช้ในการพัฒนาและวางแผนการแข่งขัน

Prescriptive Analytics แนะนำกลยุทธ์การตลาดเพื่อปรับปรุงประสิทธิภาพของทีมเบสบอล เช่น เซ็นสัญญากับผู้เล่นใหม่, พัฒนาโปรแกรมการฝึกซ้อม, วางแผนงบประมาณ เป็นต้น

ยกตัวอย่างเช่น จากการวิเคราะห์ข้อมูลสถิติของทีมและนักกีฬาเบสบอล พบว่าผู้เล่นมีอัตราการตีลูกต่ำ ซึ่งนำไปช่วยในการตัดสินใจเซ็นสัญญากับผู้เล่นที่มีทักษะการตีลูกที่ดีกว่า เป็นต้น เพื่อหาแนวทางในการปรับปรุงประสิทธิภาพของทีมให้ดียิ่งขึ้น

บทที่ 4 ผลการดำเนินงานโครงงาน

4.1. Feature Importance of Pitcher

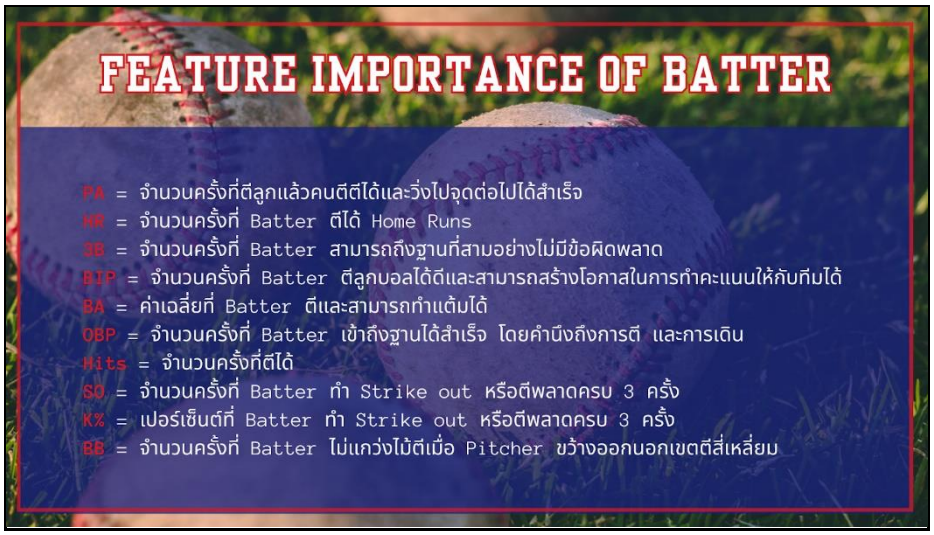


รูปที่ 4-1 รายละเอียด Feature Importance of Pitcher

จาก Feature Importance ของ Pitcher ที่มีผลต่อ Target (Pitches) โดยอนุมานว่า ยิ่งมีจำนวนขว้างที่เยอะ นั่นหมายถึง ผู้เล่นมีโอกาสในการลงแข่งเยอะ หรือก็คือสามารถทำคะแนนได้ มีความสอดคล้องกับหลักความเป็นจริงและนำไปใช้ประโยชน์ได้

ยกตัวอย่างเช่น Pitcher ที่มีจำนวน PA สูง หมายความว่า ขว้างแล้ว Batter ตีได้และสามารถทำคะแนนได้สำเร็จ ดังนั้น ค่า PA จึงส่งผลต่อจำนวนการขว้างและค่า PA ควรมีจำนวนที่ต่ำ เพราะยิ่งต่ำ ก็หมายความว่า Batter ไม่สามารถตีได้ และ Pitcher ก็จะมีโอกาสในการทำคะแนนมากขึ้น หรืออีกตัวอย่าง SO จำนวนที่ขว้างแล้ว Batter ตีพลาด 3 ครั้ง (Strikeout) โดยยิ่ง SO เยอะเท่าไหร่ ก็จะทำให้ Pitcher มีโอกาสในการทำคะแนนได้มากขึ้นเท่านั้น ซึ่งก็จะนำไปสู่การพัฒนาทักษะว่า ถ้าต้องการให้ Batter ตีพลาด ก็ควรไปฝึกเทคนิคต่าง ๆ เพิ่ม เช่น การขว้างแบบหมุน (Spin) เป็นต้น

4.2. Feature Importance of Batter



รูปที่ 4-2 รายละเอียด Feature Importance of Batter

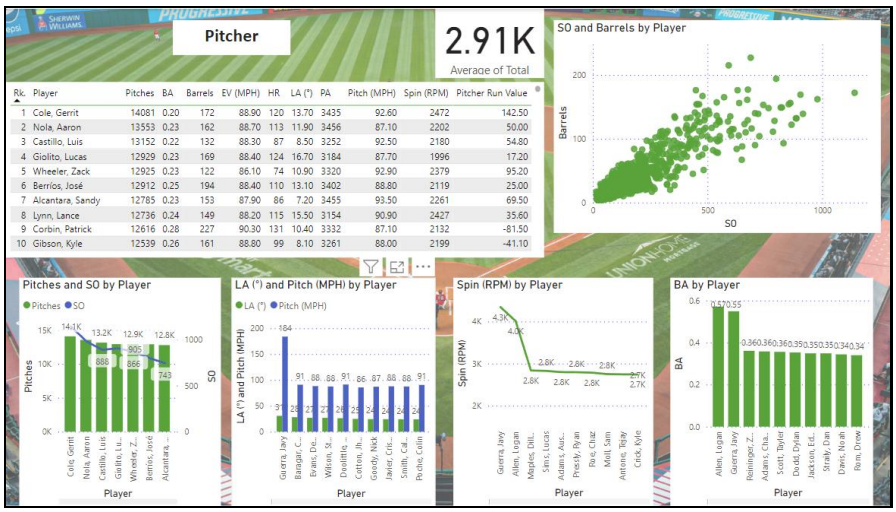
จาก Feature Importance ของ Batter พบว่ามีความสอดคล้องกับความเป็นจริงและนำไปใช้ประโยชน์ในด้านต่าง ๆ ได้

ตัวอย่างเช่น Batter ที่มีค่า HR สูง หรือก็คือ มีจำนวนครั้งที่ตี Home Runs ได้เยอะ ก็จะมีแนวโน้มที่จะทำให้โอกาสในการทำคะแนนเพิ่มขึ้น ซึ่งจะส่งผลต่อประสิทธิภาพของ Batter ในทางบวก และสามารถนำไปใช้ประโยชน์เกี่ยวกับการประเมินศักยภาพหรือพัฒนากลยุทธ์การฝึกซ้อมได้

หรือ Batter ที่มี Feature Importance สูงในปัจจัย Hits และ Home Runs แต่มี Feature Importance ต่ำในปัจจัย OBP หมายความว่า Batter คนนั้นมีแนวโน้มที่จะตีโดนบอลบ่อย ตีโฮมรันได้บ่อย แต่มีโอกาสเข้าถึงฐานได้ไม่บ่อย หาก Batter ต้องการบรรลุเป้าหมาย Pitches ให้ได้มากที่สุด Batter จำเป็นต้องปรับปรุงทักษะ โดยอาจฝึกฝนเทคนิคต่าง ๆ เช่น การเลือกตำแหน่งยืน, ทักษะในการวิ่ง เป็นต้น

4.3. Dashboard

มีทำการ Dashboard ที่แสดงผลข้อมูลที่ได้จากการวิเคราะห์ โดยใช้เครื่องมือ Power BI เพื่อสามารถนำไปใช้ประโยชน์ได้มากขึ้น



Rk.	Player	Pitches	BA	Barrels	EV (MPH)	HR	LA (°)	PA	Pitch (MPH)	Spin (RPM)	Pitcher Run Value
1	Cole, Gerrit	14081	0.20	172	88.90	120	13.70	3435	92.60	2472	142.50
2	Nola, Aaron	13553	0.23	162	88.70	113	11.90	3456	87.10	2202	50.00
3	Castillo, Luis	13152	0.22	132	88.30	87	8.50	3252	92.50	2180	54.80
4	Giolito, Lucas	12929	0.23	169	88.40	124	16.70	3184	87.70	1996	17.20
5	Wheeler, Zack	12925	0.23	122	86.10	74	10.90	3320	92.90	2379	95.20
6	Berríos, José	12912	0.25	194	88.40	110	13.10	3402	88.80	2119	25.00
7	Alcantara, Sandy	12785	0.23	153	87.90	86	7.20	3455	93.50	2261	69.50
8	Lynn, Lance	12736	0.24	149	88.20	115	15.50	3154	90.90	2427	35.60
9	Corbin, Patrick	12616	0.28	227	90.30	131	10.40	3332	87.10	2132	-81.50
10	Gibson, Kyle	12539	0.26	161	88.80	99	8.10	3261	88.00	2199	-41.10

รูปที่ 4-3 Dashboard ของ Pitcher

จาก Dashboard ของชุดข้อมูล Pitcher จะแสดงให้เห็นถึง จำนวนเฉลี่ยของเกมที่ผู้เล่นลง ซึ่งจะอยู่ที่ 2.91 K ต่อคน และจะแสดง 10 อันดับผู้เล่นเรียงตาม Pitches (จำนวนครั้งที่ขว้างได้)

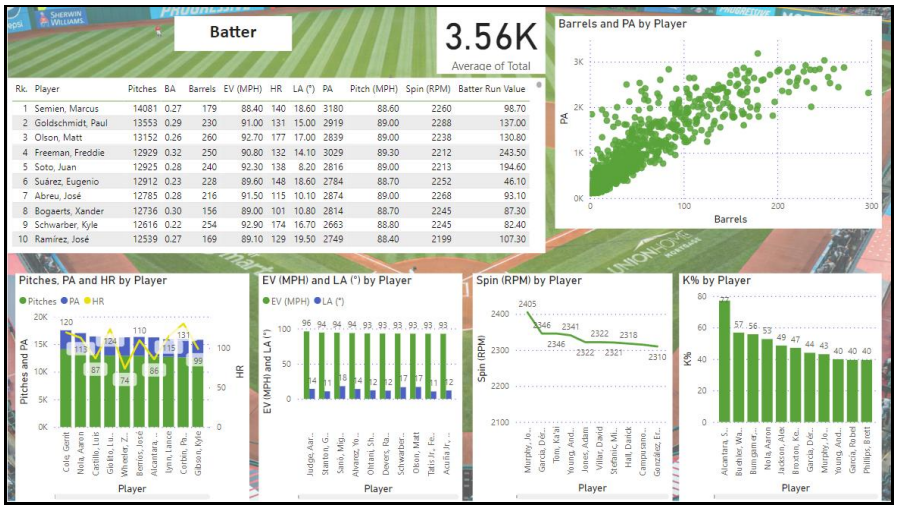
จากกราฟ Pitches and SO by Player โดย SO คือ จำนวนที่ทำ Strikeout ได้ กล่าวคือ ผู้เล่นที่มี Pitches สูง หรือ จำนวนครั้งที่ขว้างสูง ก็จะมีค่า SO ที่สูงตาม และถ้าค่า Pitches ต่ำ ค่า SO ก็จะต่ำเช่นเดียวกัน ซึ่งก็จะหมายความว่า ผู้เล่นที่สามารถทำ Strikeout ได้ก็จะมีโอกาสได้ลงสนามมากขึ้นเนื่องจากสามารถทำคะแนนได้

และจากความสัมพันธ์ของ Barrels (คนที่ขว้างได้ดีที่สุดทั้งองศาและความเร็วในการขว้าง) and SO ในรูปแบบของกราฟ Scatter Plot จะเห็นได้ว่ามีความสัมพันธ์ในเชิงบวกต่อกัน ซึ่งจะนำไปสู่กราฟ LA (°) and Pitch(MPH) เพื่อดูว่าผู้เล่น 10 อันดับแรกนั้นมีความเร็วและองศาโดยเฉลี่ยแล้วประมาณเท่าไหร่ โดย LA (°) คือ องศาในการขว้างและ Pitch(MPH) คือ ความเร็วในการขว้าง

ซึ่งจะมีองศาในการขว้างเฉลี่ยอยู่ที่ประมาณ 12 องศา และความเร็วเฉลี่ยที่ 90 MPH ซึ่งจะเป็นองศาและความเร็วในการขว้างที่จะทำให้เกิดการ Strikeout หรือ ทำให้ Batter พลาดได้มากที่สุด

ในส่วนของกราฟ Spin(RPM) by Player จะเห็นได้จาก 2 คนแรกว่า Pitcher ที่ขว้างลูกแบบ Spin (หมุน) มากก็จะเพิ่มโอกาสในการทำแต้มได้มากขึ้นถึงประมาณ 20 เปอร์เซ็นต์ ซึ่งจะสอดคล้องกับกราฟ BA by Player ที่ว่า ยิ่งเราขว้างรูปแบบ Spin ก็จะทำให้ BA (เปอร์เซ็นต์ในการขว้างลูกได้ของ Pitcher) หรือการทำคะแนนได้เยอะมากขึ้น

ในส่วนของข้อมูลชุด Batter สามารถนำข้อมูลที่ได้จากการวิเคราะห์มาแสดงผลได้ ดังนี้



รูปที่ 4-4 Dashboard ของ Batter

จากกราฟ Table จะแสดงถึง 10 อันดับผู้เล่น Batter เรียงตาม Pitches (จำนวนครั้งที่ตีได้) และจำนวนเฉลี่ยของเกมที่ผู้เล่นลงต่อคนจะอยู่ที่ประมาณ 3.56 K

กราฟ Pitches, PA and HR by Player โดย PA คือ จำนวนครั้งที่ Batter มีโอกาสตี และ HR คือ จำนวนที่ทำ Home Run ได้ ซึ่งสามารถแสดงให้เห็นถึงความสัมพันธ์กับ Pitches ว่ามีความสอดคล้องกันในเชิงบวก จากนั้นทำการแสดงความสัมพันธ์ของ PA กับ Barrel (คนที่ตีได้ดีที่สุดทั้งองศาและความเร็วในการตี) ในรูปแบบ Scatter Plot จะเห็นได้ว่ามีแนวโน้มความสัมพันธ์ไปในทิศทางเดียวกัน

จากกราฟ EV(MPH) and LA (°) by Player จะได้ว่า 10 อันดับผู้เล่นเรียงตาม Pitches โดยเฉลี่ยจะมีองศาในการตีอยู่ที่ประมาณ 15 องศา และความเร็วของลูกหลังที่ตีออกไปเฉลี่ยจะอยู่ที่ประมาณ 94 MPH ซึ่งจะเป็นค่าเฉลี่ยในการตีที่สามารถทำให้คู่แข่งมีโอกาสพลาด

ในส่วนของกราฟ Spin (RPM) คือการตีลูกให้หมุน จะเห็นได้ว่า การตีลูกให้หมุนไม่ได้ส่งผลต่อโอกาสในการทำคะแนนมากเท่าไหร่ เมื่อเทียบกับฝั่ง Pitcher ที่ยิ่งสามารถขว้างลูกแบบหมุนก็จะมีโอกาสในการทำคะแนนมากขึ้น

และจากกราฟ K% (เปอร์เซ็นต์การตีพลาด 3 ครั้ง หรือ Strikeout) จะสามารถสรุปได้ว่า ผู้เล่น 10 อันดับเรียงตาม Pitches มีค่า K% ในระดับที่ต่ำกว่า 30% ซึ่งหมายความว่า จำนวนการ Strikeout มีความสำคัญและควรคำนึงถึงต่อการทำคะแนน ยิ่งตีพลาดในระดับที่ต่ำก็จะสามารถมีโอกาสในการทำคะแนนหรือชนะมากขึ้น

ดังนั้นจะสามารถสรุปได้ว่า ผลการดำเนินงานของโครงงาน สามารถทำให้ทั้ง Pitcher และ Batter พัฒนาทักษะที่จะทำให้มีโอกาสในการชนะมากขึ้น และผู้บริหารจัดการทีมสามารถนำไปใช้ในการวางแผนกลยุทธ์ หรือจัดสรรงบประมาณ เพื่อให้เกิดประโยชน์ให้สูงที่สุดและสร้างกำไรได้เพิ่มมากขึ้น

บทที่ 5 ความเชื่อมโยงกับวิชาต่าง ๆ ในโมดูล

5.1. วข.310 การสำรวจและการเตรียมข้อมูล

ประยุกต์ใช้วิชาวข.310 การสำรวจและการเตรียมข้อมูล ในขั้นตอนของการรวบรวมข้อมูล ในการทำ Web scraping โดยใช้ BeautifulSoup และ Selenium รวบรวมข้อมูลต่าง ๆ จากเว็บไซต์ ซึ่งจะได้ออกมาดังนี้

1. ข้อมูลของชุด Pitcher ประกอบด้วย 1047 ข้อมูล

Rk.	Player	Pitches	Total	Pitch %	PA	BIP	Hits	3B	HR	SO	...	Downward Movement w/ Gravity (in)	Glove/Arm-Side Movement (in)	Vertical Movement w/o Gravity (in)	Movement Toward/Away from Batter (in)	EV (MPH)	LA (°)	Dist (ft)	Hard Hit%	Barr
1	Cole, Gerrit RHP	14081	14089	99.9	3435	2073	653	5	120	1141	...	23.4	4.3	9.1	1.4	88.9	13.7	170	39.1	
2	Nola, Aaron RHP	13553	13577	99.8	3456	2236	736	12	113	979	...	33.0	5.4	3.8	0.1	88.7	11.9	145	37.2	
3	Castillo, Luis RHP	13152	13188	99.7	3252	2061	652	14	87	888	...	26.6	12.3	6.0	1.4	88.3	8.5	145	37.9	
4	Giolito, Lucas RHP	12929	12941	99.9	3184	1994	657	18	124	905	...	23.6	5.4	12.3	0.6	88.4	16.7	167	37.6	
5	Wheeler, Zack RHP	12925	12955	99.8	3320	2243	722	12	74	866	...	22.8	5.0	9.1	1.1	86.1	10.9	151	33.3	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1043	Sousa, Bennett LHP	505	505	100.0	129	95	31	0	4	22	...	24.4	2.4	9.3	0.8	89.6	8.0	155	45.3	
1044	Gose, Anthony LHP	503	504	99.8	114	61	17	0	4	37	...	20.8	3.3	10.8	1.4	91.3	19.1	200	41.0	
1045	Rodriguez, Manuel RHP	503	503	100.0	139	94	28	0	4	24	...	25.7	6.4	5.6	0.6	83.7	6.2	128	31.9	
1046	Yamaguchi, Shun RHP	501	502	99.8	119	74	28	2	6	26	...	29.2	7.7	7.9	0.1	89.1	12.5	159	37.8	
1047	Grotz, Zac RHP	501	501	100.0	118	73	25	0	4	22	...	32.1	8.6	4.1	3.2	88.8	6.1	128	37.0	

1047 rows × 37 columns

รูปที่ 5-1 ชุดข้อมูลของ Pitcher

2. ข้อมูลของชุด Batter ประกอบด้วย 804 ข้อมูล

Rk.	Player	Pitches	Total	Pitch %	PA	BIP	Hits	3B	HR	SO	...	Downward Movement w/ Gravity (in)	Glove/Arm-Side Movement (in)	Vertical Movement w/o Gravity (in)	Movement Toward/Away from Batter (in)	EV (MPH)	LA (°)	Dist (ft)	Hard Hit%	Barr
1	Semien, Marcus	14081	14089	100.0	3180	2337	755	19	140	528	...	27.9	3.7	7.7	0.6	88.4	18.6	167	37.0	
2	Goldschmidt, Paul	13553	13577	100.0	2919	1914	727	3	131	647	...	28.0	3.4	7.2	1.6	91.0	15.0	173	47.2	
3	Olson, Matt	13152	13188	99.9	2839	1828	637	4	177	664	...	27.4	4.5	7.8	2.2	92.7	17.0	175	51.1	
4	Freeman, Freddie	12929	12941	100.0	3029	2182	839	9	132	494	...	27.2	5.0	7.6	2.8	90.8	14.1	172	45.4	
5	Soto, Juan	12925	12955	99.9	2816	1819	642	10	138	478	...	28.1	5.0	7.1	2.7	92.3	8.2	161	51.2	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
800	Williams, Nick	1020	1021	100.0	125	69	16	0	2	47	...	27.4	5.2	8.0	4.5	86.7	11.2	160	30.4	
801	Gennett, Scooter	1019	1027	99.8	138	93	30	0	2	41	...	28.9	5.0	7.1	3.1	86.3	13.1	177	25.8	
802	Butler, Lawrence	1019	1020	99.8	128	88	26	0	4	35	...	28.9	5.2	6.3	4.7	88.3	16.3	154	37.5	
803	Fried, Max	1017	1019	97.0	117	75	26	0	0	33	...	26.1	5.9	8.7	2.9	88.3	-2.1	145	37.3	
804	Alcantara, Sandy	1016	1016	96.4	109	19	6	0	0	84	...	25.9	4.7	8.8	1.0	81.4	-20.9	127	21.1	

804 rows × 37 columns

รูปที่ 5-2 ชุดข้อมูลของ Batter

และมีการประยุกต์ใช้ในขั้นตอนการทำ Data Preprocessing เพื่อเตรียมความพร้อมข้อมูลในการนำไปวิเคราะห์

5.2. วข.311 อัลกอริทึมของวิทยาศาสตร์ข้อมูล

นำความรู้ในวิชา วข.311 อัลกอริทึมของวิทยาศาสตร์ข้อมูล มาประยุกต์ใช้ในการทำ Regression Model ซึ่งนั่นก็คือ LightGBMRegressor

จากนั้นได้นำไปใช้ในขั้นตอนการทำ Model Selection โดยใช้เทคนิค K-fold Cross - Validation ในการแบ่ง Validation Set เพื่อใช้ในการปรับปรุงประสิทธิภาพของโมเดลให้เหมาะสมกับข้อมูลมากที่สุด

แล้วใช้ในการทำ Feature Selection เพื่อหา Feature ที่อธิบายความความสัมพันธ์ระหว่าง Input และ Output ได้ดีที่สุด แล้วนำไปใช้ในการทำนาย โดยใช้เทคนิค Recursive Feature Elimination (RFE) ของ Wrapper Methods เป็นการลดจำนวน Feature ที่ไม่จำเป็นออกไปและเลือกลักษณะที่มีผลต่อการทำนายอย่างมีประสิทธิภาพ ซึ่งก็จะทำให้ทราบ Feature Importance ของข้อมูล และใช้ค่า Mean Absolute Error (MAE) ในการวัดประสิทธิภาพของโมเดล เนื่องจากค่า MAE เหมาะสมสำหรับข้อมูลที่ทราบ Outliers

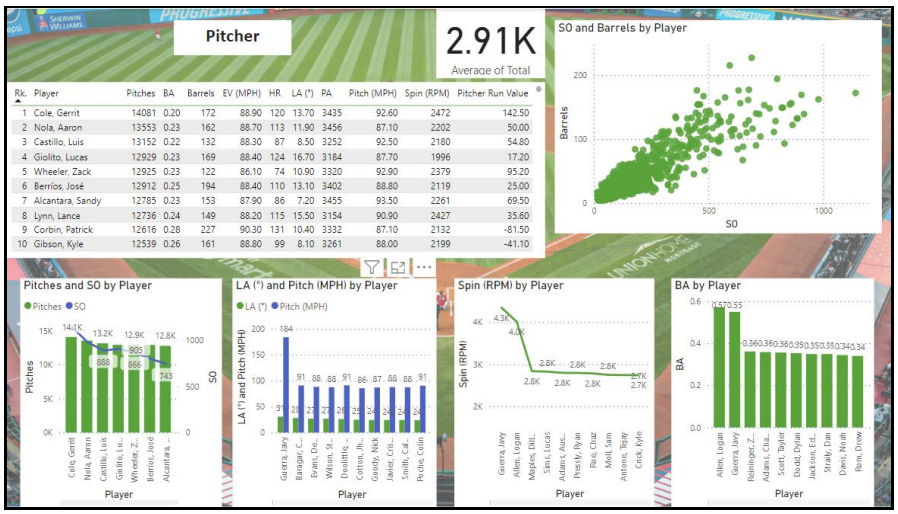
โดยมีสูตรทางคณิตศาสตร์ ดังนี้

$$\mathrm{MAE}(\mathbf{X}, h) = \frac{1}{m}\sum_{i=1}^{m}\left|h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right|$$

รูปที่ 5-3 สูตรคำนวณ MAE

5.3. วข.312 ระบบธุรกิจอัจฉริยะ

นำวิชาวข.312 ระบบธุรกิจอัจฉริยะ มาประยุกต์ใช้ในขั้นตอนการทำ Dashboard เพื่อแสดงภาพรวมของข้อมูล เพื่อให้มีความเข้าใจภาพรวมของข้อมูลมากขึ้น และสามารถนำไปตอบโจทย์กับคำถามทางการตลาด รวมถึงสามารถนำข้อมูลไปทำการวิเคราะห์ต่อได้ เพื่อนำไปใช้ประโยชน์และสร้างคุณค่าอื่น ๆ เพิ่มเติม โดยผ่านเครื่องมือ Power BI



รูปที่ 5-4 ตัวอย่าง Dashboard

5.4. วข.313 การวิเคราะห์การตลาด

ได้นำความรู้ในวิชา วข.313 การวิเคราะห์การตลาด ไปประยุกต์ใช้กับการวิเคราะห์การตลาด และมีการใช้เครื่องมือทางการตลาด โดยการตั้งคำถามตามหลัก SMART เพื่อกำหนดเป้าหมายที่ชัดเจนและสามารถนำไปสู่การบรรลุเป้าหมายได้ และมีการใช้ Market Analytics เพื่อให้เกิดความเข้าใจในข้อมูลรวมถึงสาเหตุที่เกิดขึ้นและสามารถนำไปใช้ในการเตรียมพร้อมรับมือกับเหตุการณ์ที่อาจเกิดขึ้น และช่วยให้สามารถใช้ในการกำหนดแนวทางการแก้ปัญหาและปรับปรุงประสิทธิภาพ เพื่อช่วยให้ตัดสินใจได้อย่างเหมาะสมและมีประสิทธิภาพมากขึ้น

บทที่ 6 บทสรุป

การวิเคราะห์กีฬาเบสบอลโดยใช้ข้อมูลสถิติจากการแข่งขัน MLB เพื่อใช้ในการหาปัจจัยที่มีผลต่อโอกาสในการชนะ ซึ่งจะใช้ LightGBMRegressor สำหรับการสร้างโมเดล โดยผลลัพธ์ที่ได้มีประสิทธิภาพและช่วยให้ทราบถึง Feature Importance ที่มีผลต่อโอกาสในการชนะได้อย่างชัดเจน

การวิเคราะห์ที่ดีและการสร้างโมเดลที่มีประสิทธิภาพสามารถช่วยให้องค์กรกีฬาสามารถพัฒนากลยุทธ์ทางธุรกิจในการสร้างรายได้และเพิ่มมูลค่าได้อย่างมีประสิทธิภาพ โดยการนำข้อมูลจากการวิเคราะห์ Feature Importance ที่บ่งบอกถึงปัจจัยที่มีผลต่อโอกาสในการชนะในการแข่งขันไปใช้ในการพัฒนากลยุทธ์ทางธุรกิจของผู้บริหารจัดการทีมหรือผู้ประกอบการ ซึ่งจะทำให้ทีมกีฬาสามารถเสริมสร้างชื่อเสียงและเพิ่มโอกาสในการได้รับการสนับสนุนจากสปอนเซอร์ (Sponsorship), การขายบัตรเข้าชม (Ticket Sales) และการขายสินค้าที่เกี่ยวข้อง (Merchandise) แล้วสามารถสร้างรายได้และเพิ่มมูลค่าให้กับทีมกีฬาได้อย่างมีประสิทธิภาพ

ดังนั้นการนำ Feature Importance มาปรับปรุงกลยุทธ์การตลาดและกลยุทธ์ทางธุรกิจของทีมกีฬาจะช่วยเพิ่มโอกาสในการสร้างรายได้และเสริมมูลค่าในระยะยาว และจะเป็นเครื่องมือที่มีประสิทธิภาพในการช่วยตัดสินใจทางธุรกิจของผู้บริหารจัดการทีมหรือผู้ประกอบการ

บรรณานุกรม

กีฬาทีมเบสบอล. (2561). *กติกาเบสบอล.* สืบค้นจาก https://sportthailandtoday.wordpress.com/tag/กติกาเบสบอล/

อิสรา สุนทรวัฒน์. (2566). BASEBALL จะครองโลก!!!!. สืบค้นจาก https://www.thaipost.net/columnist-people/348742/

Apipoj Piasak. (2020). *Scale or Standardize or Normalize*. Retrieved from https://medium.com/data-espresso/scale-or-standardize-or-normalize-ef905613f275

Baseballmania. (2023). *กีฬาเบสบอล*. สืบค้นจาก https://baseballmania.info/

Christoph Hoog Antink, Anne K Braczynski, Bergita Ganse. (2021). *Learning from machine learning: prediction of age-related athletic performance decline trajectories.* Retrieved from https://pubmed.ncbi.nlm.nih.gov/34241807/

Jason Brownlee. (2020). *How to Scale Data With Outliers for Machine Learning*. Retrieved from https://machinelearningmastery.com/robust-scaler-transforms-for-machine-learning/

Key.pettakon. (2019). *Feature selection 101*. สืบค้นจาก https://medium.com/@key.sompornpettakon.statkmitl/feature-selection-101-9eb8cf362dff

Mitchell F Aarons, Chris M Young, Lyndell Bruce, Dan B Dwyer. (2023). *Real time prediction of match outcomes in Australian football.* Retrieved from https://pubmed.ncbi.nlm.nih.gov/37733399/

MLB Advanced Media, LP. (2023). *Savant*. Retrieved from https://baseballsavant.mlb.com/

Paeng @DATACUBATOR. (2021). *Validation set สำคัญไฉน?*. สืบค้นจาก https://medium.com/datacubator/validation-set-สำคัญไฉน-1abf22a68b75

Sasiwut Chaiyadecha. (2021). *สร้าง Machine learning model ด้วย Pipeline*. สืบค้นจาก https://lengyi.medium.com/scikit-learn-model-pipeline-4c155228f184