דוח הכנת הנתונים

ניהול צריכה באמצעות סריקת חשבוניות קבלה

נופר גרשוני-322773680

עלית בן חמו-206851081

למנחה-גב' קרן סגל

תוכן עניינים:

1. בחירת נתונים:

בהתבסס על איסוף הנתונים שבוצע בשלב הקודם, שלב בחירת הנתונים מתמקד בבחירת פריטים ומאפיינים הרלוונטיים למטרת חיזוי אי ההתאמות בכמויות המוצרים.

1.1. בחירת פריטים-

החלטנו להמשיך ולאסוף חשבוניות נוספות כדי להרחיב את בסיס הנתונים ולהבטיח כמות נתונים גדולה ככל האפשר. הרחבה זו תסייע בזיהוי מדויק יותר של דפוסים והתנהגויות חריגות, ובכך תשפר את יכולות החיזוי והדיוק, דבר המוביל לתחזיות אמינות יותר.

1.2.בחירת מאפיינים-

בחירת המאפיינים מתייחס לערכים שמספקים את המידע הקריטי לחיזוי אי ההתאמות. המאפיינים הנבחרים כוללים:

מק"ט: מזהה ייחודי של המוצר.

שם המוצר: תיאור המוצר לצורך זיהוי מדויק.

כמות: המאפיין המרכזי המייצג את מספר הפריטים בכל חשבונית.

מחיר: מאפשר בדיקה והשוואה בין כמות המחיר לרכישות.

מספר חשבונית: מאפשר זיהוי ובדיקה של אי התאמות בכל חשבונית.

מספר לקוח: מספק נתון חשוב לניתוח היסטוריית הרכישות ולחזות אי התאמות בהתאם לתדירות הרכישות של כל לקוח.

1.3.שיקולים ותיעוד החלטות-

במהלך בחירת הנתונים נשקלו מספר שאלות מרכזיות:

רלוונטיות המאפיינים: כל המאפיינים הנבחרים תורמים באופן ישיר למטרת החיזוי.

איכות הנתונים: על מנת להבטיח את אמינות התוצאות, בוצעו תהליכי ניקוי ושיפור איכות הנתונים. אחת הפעולות המרכזיות הייתה הסרת כפילויות, כדי למנוע השפעה של רשומות חוזרות שעלולות להטות את הממצאים ולפגוע בדיוק החיזויים. בנוסף, בוצע איחוד של מוצרים דומים, כך מוצרים בעלי שמות קרובים או הבדלים קלים בכתיב לא יופיעו כרשומות נפרדות, מה שמאפשר לזהות מגמות צריכה באופן מדויק יותר.

כבר בשלב הבנת הנתונים חושבו סטיית התקן והכמות הממוצעת של המוצרים, כדי לזהות את החריגות בכמות. זיהוי חריגות המתבסס על סטיות משמעותיות בהתנהגות הצרכנית הממוצעת, דבר שאפשר לשפר את הביצועים של המערכת התחזיתית.

בנוסף, לא נמצאו נתונים חסרים, וכל הנתונים שנבחרו הם מלאים ועקביים, מה שמבטיח שהמערכת אינה נדרשת לבצע שחזור או השלמה של מידע חסר. מצב זה משפר את רמת האמינות של התוצאות ומאפשר עיבוד נתונים בצורה יציבה ומדויקת.

מגבלות שימוש: יש לשמור על הפרדה בין המאפיינים העיקריים לבין אלו שאינם תורמים ישירות לתהליך החיזוי, ולהקפיד על שמירת פרטיות המידע, במיוחד בכל הנוגע למזהי לקוחות.

1.4.התאמה להנחות מהשלב הקודם-

הבחירה של הנתונים בשלב זה מתבססת במלואה על ההנחות והבדיקות שבוצעו בשלב הבנת הנתונים. במסמך הקודם חישבנו את הכמות הממוצעת ואת סטיית התקן של המוצרים עבור כל מספר לקוח, ובהתבסס על כך קבענו גבולות עליונים ותחתונים לזיהוי אי התאמה, באמצעות שתי סטיות תקן מהממוצע. הגדרת גבולות אלו מאפשרת לזהות מקרים שבהם הכמות חורגת מהתנהגות הרכישה הצפויה של כל לקוח, ובכך מספקת בסיס מדעי לאיתור אי התאמות. בנוסף, על מנת לשפר את היכולת לזהות חריגות בהתנהגות הצרכנית, הוספנו לנתונים את העמודות המחושבות של סטיית תקן, כמות ממוצעת, גבול תחתון וגבול עליון.

במהלך חקר הנתונים, ההשערות שהועלו אושרו כאשר נמצא כי אכן קיימות אי התאמות בכמויות המוצרים בהיסטוריית רכישות. בנוסף, זוהו מאפיינים מרכזיים אשר נמצאו כבעלי תרומה משמעותית להמשך הניתוח, בהם הכמות של המוצרים ומספר הלקוח. מאפיינים אלו מספקים תשתית חזקה לדיוק החיזוי והיכולת של המערכת לזהות חריגות רלוונטיות.

הניתוח הוכיח כי לרוב הלקוחות קיימים אי-התאמות בכמות המוצרים, מה שהדגיש את הצורך בהתמקדות בניתוחים מותאמים אישית לכל לקוח. בהתאם לכך, בוצעה בדיקה פרטנית לכל לקוח, שבה נבדקה כל רכישה מול הטווחים המחושבים, וכך ניתן היה לאמת את קיומן של אי התאמות באופן שיטתי.

ממצאים אלו מחזקים את תקפות ההנחות הראשוניות, ומאפשרים המשכיות בתהליך בניית המודלים.

2. ניקוי נתונים:

תהליך ניקוי הנתונים כלל בחינה מעמיקה של הבעיות שהופיעו בנתונים שנבחרו לניתוח, במטרה להסיר או לתקן את "הרעש" ולהבטיח שהנתונים יהיו עקביים ואמינים להמשך ניתוח וחיזוי. להלן פירוט הסוגים שהתגלו והפעולות שננקטו:

2.1.סוגי רעש שהתגלו-

התגלה רעש בנתונים עקב כפילויות ושגיאות כתיב בשמות המוצרים. במקרים מסוימים, שמות מוצרים זהים הופיעו מספר פעמים בתוך אותה חשבונית, כתוצאה משגיאות כתיב או וריאציות שונות של שם המוצר. כדי להתמודד עם בעיה זו, איחדנו את המוצרים בעלי שמות דומים, כך שכל מוצר מוצג פעם אחת בלבד, מה שמונע עיוותים בנתונים ומשפר את דיוק הניתוחים.

2.2.שיטות להסרת הרעש וטכניקות מוצלחות-

איחוד רשומות כפולות: בוצעה בדיקה של רשומות עם שמות מוצרים דומים, והוחלט על איחודם לשם הבאת אחידות במידע.

המרת פורמטים ובחירה סכמת קידוד אחידה: כל הערכים הקטלוגיים עברו המרה לפורמט תקני על מנת להבטיח אחידות בקידוד ובייצוג הנתונים.

בדיקות לוגיות: נערכו בדיקות לוגיות על הנתונים כדי לוודא את תקינותם ודיוקם. תחילה, בוצעה בדיקה לווידוא שכל הערכים בעמודות הכמות והמחיר הם מספריים ואינם מכילים ערכים שליליים. לאחר מכן, חושבו הממוצע וסטיית התקן של כמות הרכישות עבור כל לקוח, ונקבעו טווחים תחתון ועליון המבוססים על שתי סטיות תקן מהממוצע. כל ערך שחורג מהטווח שנקבע סומן כחריגה. זיהוי חריגות התבצע באופן אוטומטי על ידי השוואת הכמות בכל עסקה לטווח הרלוונטי של אותו לקוח. בנוסף, לכל ערך שזוהה כחריג חושבה מידת החריגה ביחס לממוצע כדי להעריך את סטייתו מההתנהגות הצפויה של הלקוח. תהליך זה בוצע בפייתון.

3. יצירת נתונים חדשים:

יצירת נתונים חדשים הוא שלב חשוב בתהליך שיפור נתונים לצורך מודלים לחיזוי. ישנן שתי דרכים עיקריות בהן ניתן לפעול:

3.1.הפקת מאפיינים חדשים-

נרמול נתונים: מאחר שיש לנו מאפיינים מספריים כמו כמות ומחיר,ביצענו נרמול כדי להבטיח שהמודלים לא יושפעו ממאפיינים בעלי סקאלה שונה. לדוגמה, נרמול כמות המוצרים ומחירים יכול לשפר את ביצועי המודל ולהפוך את הנתונים לאחידים ומותאמים למודלים שונים.

חישוב מאפיינים חדשים: הכמות הממוצעת של כל מוצר לכל לקוח, את סטיית התקן, את הגבולות התחתונים והעליונים על פי סטיית התקן, על מנת לזהות את החריגות, בנוסף הוספנו מאפיינים סטטיסטיים (סטיית תקן וממוצע) לכל לקוח ולכל מוצר(מק״ט) של כל לקוח, הפרש בין הכמות המקסימלית והמינימלית של הלקוח ואת ההפרש בין המחיר המקסימלי והמינימלי של הלקוח.

3.2.יצירת רשומות חדשות-

ניתן לבצע אגרגציה על הנתונים כדי להוסיף את המאפיינים המחושבים כרשומות חדשות. רשומות אלו עשויות לשפר את ניתוח החריגות, ולספק תובנות נוספות שיכולות לעזור בשיפור המודל.

דרישות המודל:

בחרנו להשתמש בעמודות הקטגוריאליות "שם מוצר" ו-"שם החנות" מתוך הנתונים, תוך המרת לערכים מספריים באמצעות שיטת קידוד One-Hot Encoding, על מנת להתאים לדרישות המודלים החיזויים. למרות שעמודות אלו אינן תורמות באופן ישיר לחישוב אי-ההתאמה בכמויות, הן עשויות לספק מידע נוסף שממנו המודלים יכולים ללמוד ולשפר את יכולת החיזוי. עמודת "שם מוצר" לא הוסרה, אלא קודדה בהתאם, אף על פי שהייתה קיימת גם עמודת "מקט" שמייצגת את המוצר באופן מספרי, מתוך מטרה לאפשר למודלים להיעזר בכל מקור מידע אפשרי. בדרך זו, שמרנו על עקביות והבטחנו שהנתונים יהיו מותאמים לדרישות המודלים, תוך שמירה על מידע פוטנציאלי שיכול לשפר את ביצועי החיזוי.

הסקה לוגית: על פי הידע הקיים, ניתן להפיק עובדות חשובות מתוך שדות קיימים. לדוגמה, ניתן להשוות בין מחיר המוצר לכמות הרשומה בחשבונית. אם יש חוסר התאמה, ניתן להפיק עובדה לגבי חריגה בכמות.

4.שילוב נתונים:

קיימים מספר מקורות נתונים שיכולים להכיל מידע שונה אך קשור לאי-התאמות בכמויות. כל מקור נתונים יכול לכלול פרטים שונים, כמו:

נתוני רכישות (כוללים מידע על המוצרים, כמו כמות ומחיר).

נתוני לקוחות (היסטוריית הרכישות של כל לקוח).

נתוני חשבוניות (כוללים פרטים על כל רכישה, כמו מספר חשבונית, מחיר וכמות).

4.1.מיזוג נתונים-

השתמשנו בפעולת המיזוג כדי לשלב נתוני רכישות, נתוני לקוחות ונתוני חשבוניות עם רשומות דומות, אך כל אחד מהם מכיל מאפיינים שונים.מימשנו את זה כך על ידי שימוש במזהה ייחודי (למשל מזהה מוצר, מזהה לקוח, מזהה חשבונית) כך נקבל סט נתונים אחד שיש בו את כל הרשומות המתאימות מכל מערך הנתונים.

על ידי ריכוז נתוני רכישות, לקוחות וחשבוניות במסד נתונים אחד, הבטחנו שכל המידע הרלוונטי לאי-התאמות בכמויות יהיה נגיש בקלות ומאורגן בצורה מסודרת. ארגון זה שיפר את ניהול וניתוח הנתונים בצורה יעילה, ואיפשר לנו להפיק תובנות מדויקות ולקבל החלטות מושכלות לגבי אי-התאמות. שילוב של נתוני רכישות, לקוחות וחשבוניות במסד נתונים אחד שיפר את הנגישות, העקביות והאמינות של הנתונים, ומפשט את תהליך הניתוח והדיווח.

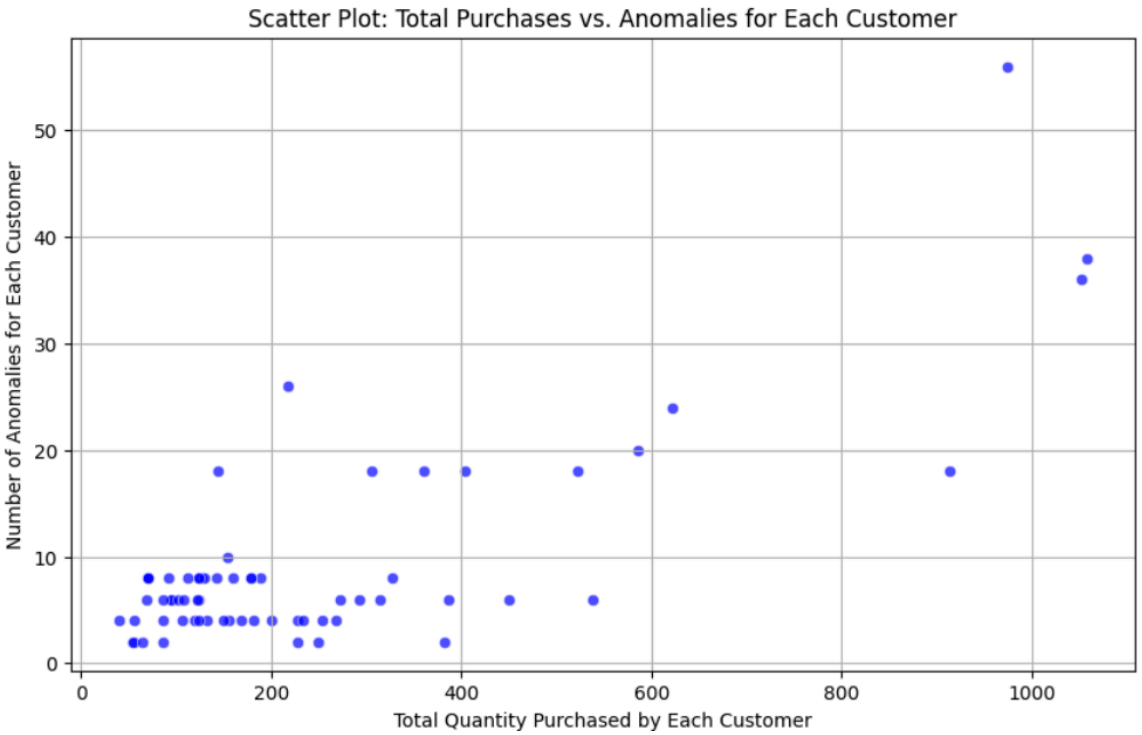
5. עיצוב נתונים:

לאחר שלב הבנת הנתונים, עברנו לשלב הכנת הנתונים, שבו ביצענו ניקוי, עיבוד והתאמה של מערכי הנתונים כך שיתאימו לניתוח זיהוי אי-התאמות בכמויות. במהלך תהליך זה, הסרנו מידע בלתי רלוונטי וביצענו אופטימיזציה של הנתונים כדי להבטיח שהם מוכנים לשימוש במודל החיזוי. בשלב הבא איחדנו נתונים רלוונטיים, כך נתוני רכישות, חשבוניות ולקוחות שולבו ליצירת סט נתונים אחיד מובנה. השתמשנו בטכניקות כמו מיזוג והצטרפות כדי לשלב נתונים ממקורות שונים וליצור מסגרת נתונים המאפשרת ניתוח מעמיק ואיתור אי-התאמות. כדי להבטיח שהמודל יכול לזהות אי-התאמות בצורה מדויקת, ביצענו את הפעולות הבאות: בחירת מודלים מתאימים, בהם מודל רגרסיה הדורש נתונים מספריים ומובנים, ארגון הנתונים בפורמט מתאים, כך שכל הנתונים יהיו מסודרים בצורה עקבית ללא מאפיינים קטגוריאליים שיכולים לפגוע בביצועים, ובדיקת תקינות והתאמות נוספות במידת הצורך. מאחר שהעמודות כמות ומחיר כבר נורמלו, נבצע בדיקות נוספות אם ימצאו פערים או חוסר אחידות בנתונים. בנוסף, תיעדנו את כל השינויים שנעשו בנתונים, כדי להבטיח עקביות ולשמור על שקיפות. בהמשך, נשתמש בכלים מתקדמים לניתוח והצגת תובנות, כך שנוכל להמיר את הנתונים הגולמיים למידע ברור ונגיש, ובאמצעות ויזואליזציה נוכל להציג אי-התאמות בצורה ברורה ולסייע למשתמשים להבין במהירות את הפערים ולהגיב בהתאם.

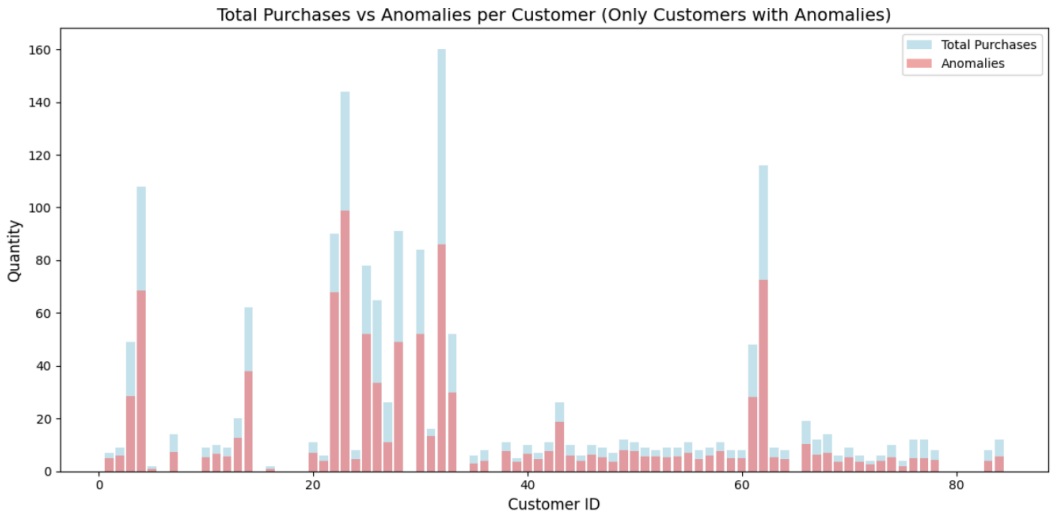
6. ניתוח נתונים EDA:

בנוסף לניתוח נתונים שעשינו בשלב הקודם, בשלב זה, אנו מעמיקים ומבצעים חקירה מעשית של הנתונים כדי לחלץ מהם תובנות חשובות ומדויקות יותר. בעזרת טכניקות וניתוח סטטיסטי, נוכל לזהות דפוסים ואי-סדרים בנתונים שיסייעו לנו להבין את התופעות באופן יותר רחב. באמצעות גישה זו, אנו שואפים להבין בצורה מעמיקה יותר את הדפוסים הקיימים בנתוני אי-ההתאמות בין רכישות שונות, מה שיסייע לנו להפיק תובנות המאפשרות למשתמשים להבין טוב יותר את בעיות אי-ההתאמה בכמויות ולהגיב בהתאם.

גרף הפיזור מציג את הקשר בין סך כמות הרכישות של כל לקוח לבין מספר אי-התאמות שהתגלו עבורו בהתאם לטווח הצפוי, בהתבסס על סטיית התקן, כמות ממוצעת וגבולות תחתונים ועליונים. כל נקודה בגרף מייצגת לקוח, כאשר ציר ה-X מציין את סך כל הרכישות שביצע הלקוח (סך הכמויות הכוללות של הפריטים שנרכשו בכל החשבוניות), וציר ה-Y מציין את מספר אי-התאמות שהתגלו עבור הלקוח. לדוגמה, נקודה בגרף מציינת לקוח שסך הרכישות שלו הוא 1000 (כלומר, הלקוח קנה 1000 פריטים בסך כל רכישותיו), ומספר החריגות שלו הוא 60, אז זה אומר שהלקוח ביצע רכישות בסך 1000, ובמהלך הרכישות הללו זוהו 60 חריגות. נראה גם מגמה בגרף, שבה ככל שכמות הרכישות של הלקוח גדלה, גם מספר החריגות נוטה לעלות. תופעה זו עשויה להצביע על כך שבסך רכישות גבוהות יש יותר סיכוי להיתקל בחריגות.



גרף מציג את כמות הרכישות של כל לקוח לצד מספר אי-התאמות שהתגלו בכל רכישה, בהתבסס על גבולות תחתון ועליון שנקבעו לפי ממוצע וסטיית תקן. מהגרף ניתן לזהות מגמה ברורה: ככל שכמות הרכישות של לקוח גדולה יותר, כך גם מספר אי-התאמות עולה. לעומת זאת, כאשר כמות הרכישות נמוכה יותר, מספר אי-התאמות קטן. לדוגמא, כאשר סך הרכישות עומד על 50, מספר אי-התאמות הוא 30, ואילו כאשר סך הרכישות מגיע ל-110, מספר אי-התאמות עולה ל-70.



גרף זה מציג את אחוז האי-התאמות בכמות הרכישה מהטווח הצפוי לפי הגבולות התחתונים והעליונים שחישבנו המתבססים על סטיית התקן והממוצע של הכמות של כל לקוח, ניתן לראות בצהוב את הלקוחות הקיימים להם אי-התאמות עד 10% ובאדום את הלקוחות שהאחוז אי-התאמה בכמות הרכישה הינו מעל 10%.