**Bike Sharing**

**Data Analysis**

**חלק ב'**

תמונה שמכילה גלגל אופניים, גלגל, רכב, רכיבה על אופניים

התיאור נוצר באופן אוטומטי

**קבוצה 37**

318720711

207992397

הנדסת תעשיה וניהול – מודלים ברגרסיה לינארית 

**תוכן עניינים**

1. **תקציר מנהלים.......................................................................................................3**
2. **עיבוד מקדים..........................................................................................................5**
   1. **הסרה של משתנים......................................................................................5**
   2. **התאמת משתנים.........................................................................................5**
   3. **משתני דמה...............................................................................................6**
   4. **משתני אינטראקציה...................................................................................7**
3. **התאמת המודל ובדיקת הנחות המודל.......................................................................10**
   1. **בחירת משתני המודל................................................................................10**
   2. **בדיקה ויזואליות של הנחות המודל..............................................................12**
   3. **בדיקה פורמלית של הנחות המודל ..............................................................14**
4. **נספחים................................................................................................................15**

1. **תקציר מנהלים:**

במסגרת הפרויקט שביצענו בחנו את כמות המשתמשים המזדמנים המגיעים לשכור אופנים בחנות. חקרנו את ההשפעות של מאפיינים שונים בעונות שונות על השכרת אופנים. במטרה לבנות מודלים שיעזרו לנו לחזות ולגלות מידע חדש ויעיל, תוך שימוש בחומר הנלמד בקורס.

בחלק ב של הפרויקט עסקנו בניתוח הנתונים כאשר מה שבדקנו היה מודל הרגרסיה הטוב והמתאים ביותר לנושא הפרויקט.

תחילה, דנו באפשרות של להסיר משתנים אשר לא משפיעים בצורה משמעותית על המודל שלנו, כלומר אין מתאם בינם לבין המשתנה המוסבר שלנו שהוא כמות המשתמשים המזדמנים. על מנת לבחון את האפשרות של הסרת המשתנים הרציפים נעזרנו במקדם המתאם של פירסון בין המשתנים המסבירים למשתנה המוסבר ולפיו החלטנו לא להסיר את המשתנים הרציפים.

בשביל למצוא מודל איכותי יותר בדקנו את האפשרות להסיר חלק מהמשתנים, לכן בדקנו עבור המשתנים הרציפים בדיקות קורלציה וPvalue- עם המשתנה המוסבר. עבור hum ראינו כי גם מתקיימת קורלציה חלשה וגם Pvalue מאד גדול לכן החלטנו להסיר משתנה זה.

על מנת לבחון את האפשרות של הסרת משתנים קטגוריאליים בחנו כמה אפשרויות הגיוניות של איחוד קטגוריות, הפיכה למשתנה בינארי ויצירת משתנה חדש. לאחר בדיקת המודל עבור כל אחת מהאופציות ראינו כי הפיכת משתנה weekday למשתנה בינארי weekend שיפר את המודל בצורה המשמעותית ביותר לכן בחרנו באופציה זו.

עבור התאמה של משתנים קטגוריאליים למודל הרגרסיה שברשותנו התאמנו עבורם משתני דמה ומשתני אינטראקציה שלהם עם המשתנים הרציפים.

ביצענו ניתוח למודל הכללי שקיבלנו, הכולל את קבוצת הבסיס, ממשתני הדמה והאינטראקציה. באמצעות אלגוריתמי שיפור ״קדימה״ ו״אחורה״ על מודל הרגרסיה שברשותנו על מנת לבחור את המשתנים הנכונים ביותר עבורו.

כך קיבלנו את המודל האידאלי לפי מדד , מהאלגוריתם ״אחורה״.

לבסוף בדקנו את קיום הנחות המודל עבור מודל הרגרסיה שברשותנו. את קיום ההנחות בחנו באמצעות תרשימים שונים בהתאם לכל הנחה ובאמצעות מבחנים סטטיסטיים אשר הציגו לנו שההנחות אכן מתקיימות במודל הרגרסיה שברשותנו.

**טבלת המשתנים מחלק א׳:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **סוג המשתנה (מוסבר/מסביר)** | **סימון ב-R** | **יחידת מידה** | **סוג המשתנה (רציף/קטגוריאלי)** | **הסבר קצר על המשתנה** |
| מסביר | instant | none | רציף | משתנה מספרי המציין את מספר הרשומה בטבלה (רשומה מייצגת יום בשנה מסוימת) |
| מסביר | season | Seasons {spring, summer, winter, fall} | קטגוריאלי | משתנה המציין את עונת השנה |
| מסביר | mnth | Months {1-12} | קטגוריאלי | משתנה המציין את החודש |
| מסביר | holiday | Boolean {yes, no} | קטגוריאלי | משתנה בינאי שמציין האם היום הוא יום חג |
| מסביר | weekday | Days {0-6} | קטגוריאלי | משתנה המציין את היום בשבוע (כאשר 0 תחילת השבוע ו-6 סוף) |
| מסביר | temp | Celsius | רציף | משתנה המציין את הטמפרטורה המנורמלת |
| מסביר | atemp | Celsius | רציף | משתנה המציין את הטמפרטורה המנורמלת המורגשת |
| מסביר | hum | Percentage | רציף | משתנה המציין את הלחות המנורמלת |
| מסביר | windspeed | Km/hour | רציף | משתנה המציין את מהירות הרוח המנורמלת |
| מסביר | registered | individuals | רציף | משתנה המציין את כמות הלקוחות הרשומים שהשכירו אופניים באותו יום |
| מוסבר | casual | individuals | רציף | משתנה המציין את כמות הלקוחות המזדמנים שהשכירו אופניים באותו יום |

1. **עיבוד מקדים:**
   1. **הסרה של משתנים:**

להלן טבלת קורלציה בין כל המשתנים הרציפים:

תמונה שמכילה טקסט, צילום מסך, גופן, מספר

התיאור נוצר באופן אוטומטי

אחרי הוצאת החריגים בחלק א׳ של הפרויקט, אנו נבחן את המשתנים הרציפים בלבד ונבדוק האם כדאי להסיר חלק מהם.

נסתכל על הקורלציה של כל המשתנים הרציפים ביחס למשתנה המוסבר ונבדוק איפה יש את המתאם הנמוך ביותר.  
נראה כי משתנה hum ו windspeed בעלי המתאם הנמוך ביותר ביחס לשאר המשתנים לכן נשקול להוריד משתנים אלה, בגלל שקורלציה לא סיבה טובה מספיק נבדוק גם בעזרת Pvalue.  
לאחר בדיקת הערכים:

|  |  |  |
| --- | --- | --- |
|  | **hum** | **windspeed** |
| **מתאם(correlation )** | -0.0294 | -0.195 |
| **Pvalue** | 0.581 | 0.0002 |

נראה כי הקשר הלינארי בין hum לבין casual משמעותית חלש יותר מאשר הקשר של windspeed ו casual ,לכן את משתנה hum נחליט למחוק , מכיוון שה Pvalue של windspeed קטן מ0.05 משתנה זה נשאיר כי כנראה מקיים קשר סיבתי כלשהו.

* 1. **התאמת משתנים:**

חשבנו על כמה אפשרויות של הגדרה מחדש של משתנים כדי לחזק את הקשר בין המשתנים למשתנה המוסבר:

* temp\_diff = atemp-temp המשתנה יהיה ההפרש של שני המשתנים, משתנה חדש זה עשוי להסביר טוב יותר את השכרת האופניים, כיוון שהוא מתייחס לא רק לטמפרטורה, אלא גם לתחושת הטמפרטורה.
* Weekend ניתן להפוך weekday למשתנה קטגוריאלי בינארי כאשר "0" מייצג ימים 1-4 ו "1" מייצג ימים 5, 6.  
  כיוון שהתנהגות השכרת האופניים בסופי שבוע כנראה שונה מבימי חול.
* ניתן לאחד עבור משתנה season את הקטגוריות "Winter" עם"fall" ו springer עם summer כיוון שתנאי מזג האוויר יחסית דומים בעונות אלו.
* ניתן להפוך את "temp" למשתנה קטגוריאלי על ידי חלוקתו לקטגוריות כמו "חם", "קר", "ממוצע".

עבור temp\_diff נראה כי המתאם שלו עם casual הרבה פחות טוב מאשר temp ו atemp לכן לא נשאיר אותו.

תמונה שמכילה טקסט, גופן, צילום מסך, מספר

התיאור נוצר באופן אוטומטי

עבור weekend נראה כי המשתנה מתאים לסיפור, הגיוני לכל מדינה ובעל מתאם יחסית גבוהה של 0.2 לכן נשתמש במשתנה זה במקום weekday .

A number of numbers on a white background

Description automatically generated

במקרה של איחוד קטגוריות במשתנה season החלטנו לא לאחד כיוון שיכול להיות שיש מאפיינים ייחודיים לעונות מסוימות והיינו רוצים לראות אותם בהמשך, בנוסף לכך כיוון שאנחנו לא יודעים מאיזה מדינה נלקח המידע לא ניתן להיות בטוח כי תנאי מזג האוויר באמת דומים בין אביב לקיץ וסתיו לחורף.

במקרה של temp והפיכתו למשתנה קטגוריאלי החלטנו לא לממש זאת כי ראשית נתוני הטמפרטורה מנורמלים אז לא ניתן לדעת איזה נתון הוא "חם" או "קר", שנית בכל מדינה ההרגשה של חם וקר היא יחסית לכן לסווג לקטגוריות חם קר כנראה יביא לאיבוד מידע.

* 1. **משתנה דמה:**

בכדי לתאר משתנים קטגוריאליים נשתמש במשתנה דמה. ניצור משתנה בינארי המקבל את הערך 0 או 1 וכך נוכל לייצג קטגוריות שונות. בכל קטגוריה נגדיר משתנה אחד אשר יהווה את קבוצת הבסיס, כלומר ערכי שאר המשתנים יהיו אפס. עבור משתנה קטגוריאלי בעל n קטגוריות נגדיר n-1 משתני דמה עבור משתנה זה. המשתנים שעבורם ניצור משתני דמה הם:

**weekend\_help:**

יהיה משתנה בינארי, האם הרשומה היא בסוף השבוע או לא.

המשתנה הנ״ל, יציין את התרומה השולית של ביצוע השכרת אופניים ביום מסוים לעומת יום רגיל במשך השבוע – עבור החותך שכבר קיים במודל.

**holiday\_help:**

יהיה משתנה בינארי, האם הרשומה היא ביום חג או לא.

המשתנה הנ״ל, יציין את התרומה השולית של ביצוע השכרת אופניים ביום חג לעומת יום רגיל – עבור החותך שכבר קיים במודל.

בנוסף, נפצל את העמודה Season, המורכבת מ-4 קטגוריות של עונות השנה, ל-3 משתני עזר. כאשר העונה fall משמשת כקבוצת הבסיס.

**summer\_help:**

יהיה משתנה בינארי, האם הרשומה היא בעונה או לא. נגזר מהמשתנה season.

**winter\_help:**

**spring\_help:**

המשתנה הנ״ל, יציין את התרומה השולית של ביצוע השכרת אופניים בעונה לעומת שאר העונות – עבור החותך שכבר קיים במודל.

* 1. **משתנה אינטראקציה:**

עוזרים לשפר את מודל הרגרסיה ולהבין השפעה של משתני עזר.

עבור כל משתנה עזר נבצע משתנה אינטראקציה שלו עם משתנה רציף:

בין סוף השבוע לבין כמות השכרות רשומות:

כדי לאמוד את התרומה השולית של היום בשבוע, נגדיר את משתני האינטראקציה:

כאשר ערכים מכילים את ערכי ה-המשתנה הנ״ל, יציין את התרומה השולית של ביצוע השכרת אופניים ביום מסוים לעומת יום רגיל בתחילת השבוע – עבור החותך שכבר קיים במודל.

חשבנו שנכון למדל כך מכיוון שבימים שישי+שבת זה ימים שבדרך כלל רוב האנשים לא עובדים בהם כך שיש יותר זמן פנוי

תמונה שמכילה טקסט, תרשים, צילום מסך, קו

התיאור נוצר באופן אוטומטי

ניתן לראות את ההשפעה של משתנה האינטראקציה על המשתנה המוסבר casual:

החותך גבוה יותר והשיפוע נמוך יותר כך שהשפעתו חלשה יותר על המשתנה המוסבר

השפעת חג על המעלות:

כדי לאמוד את התרומה השולית של החג והמעלות, נגדיר את משתנה האינטראקציה:

המשתנה הנ״ל, יציין את התרומה השולית של הטמפרטורה ביום חג על ביצוע השכרת אופניים ביום מסוים לעומת יום רגיל.

חשבנו שנכון למדל כך מכיוון שבימי חג מגיעים יותר חובבנים שאחד השיקולים שלהם הוא מזג האוויר לעומת שאר הימים שמגיעים רוכבים מקצועיים שפחות מושפעים מתנאי מזג האוויר.

תמונה שמכילה טקסט, קו, תרשים, צילום מסך

התיאור נוצר באופן אוטומטי

ניתן לראות כי אין הרבה תצפיות בחגים, אולם יחד עם זאת השיפוע כמעט זהה להשפעה של הטמפרטורה בימים רגלים.

השפעת העונה והרוח:

כדי לאמוד את התרומה השולית של השפעת הרוח לפי העונה, נגדיר את משתני האינטראקציה:

תמונה שמכילה טקסט, צילום מסך, קו, תרשים

התיאור נוצר באופן אוטומטי

ניתן לראות שיפועים הפוכים.

תמונה שמכילה טקסט, קו, צילום מסך, תרשים

התיאור נוצר באופן אוטומטי

ניתן לראות שיפועים הפוכים.

נוסיף את בהתאם:

|  |  |
| --- | --- |
| משתנה | אינדקס=i |
|  | 1 |
|  | 2 |
|  | 3 |
|  | 4 |
|  | 5 |
|  | 6 |
|  | 7 |
|  | 8 |
|  | 9 |
|  | 10 |

לכן נקבל את המודל:

1. **התאמת המודל ובדיקת הנחות המודל:**
   1. **בחירת משתני המודל:**

על מנת לבחור את משתני המודל המתאימים ביותר נשתמש באלגוריתם רגרסיה לפנים ורגרסיה לאחור למציאת המודלים הטובים ביותר ונשווה בין המודלים שנקבל לפי מדד R2adj.

**רגרסיה לפנים:**

באלגוריתם זה נתחיל במודל ללא משתנים. בכל שלב שנבצע נכניס רק משתנה אחד. המשתנה

שייכנס למודל יהיה בעל המובהקות הגבוהה ביותר ולאחר מכן נשתמש במבחן F חלקי לצורך בדיקת אילו מהמשתנים מובהקים ונכניס את המשתנה המובהק ביותר.

**רגרסיה לאחור:**

באלגוריתם זה נשתמש במודל שבתוכו נמצאים כל המשתנים ונבדוק את המובהקות של כל אחד מהם. כאשר בכל שלב נוציא את המשתנה הכי פחות מובהק.

**עבור המודל המלא:**

תמונה שמכילה טקסט, צילום מסך, גופן, מספר

התיאור נוצר באופן אוטומטי

**לאחר הרצת רגרסיה לפנים:**

תמונה שמכילה טקסט, צילום מסך, גופן, מספר

התיאור נוצר באופן אוטומטי

**לאחר הרצת רגרסיה לאחור:**

תמונה שמכילה טקסט, צילום מסך, גופן, מספר

התיאור נוצר באופן אוטומטי

שני האלגוריתמים הביאו תוצאות שונות, כיוון שסדר הבדיקות היה שונה וברגע שהגיע לצעד שבו עדיף לא לשנות כלום הם הפסיקו ולא בדקו את כל הקומבינציות האפשריות ולכן לא הייתה חפיפה בבדיקת האופציות.  
לפי מדד R2adj המודל לאחר רגרסיה לאחור הוא טוב יותר מכיוון שקיבלנו את הערך הגבוה יותר ולכן נמשיך עם מודל זה.

**המודל הסופי:**

* 1. **בדיקה ויזואלית של הנחות המודל:**

תמונה שמכילה טקסט, תרשים, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

בבדיקת **הנחת הלינאריות** נעזר בתרשים פיזור השאריות המתוקננות ונבחן את צורת הפיזור של התצפיות ונחפש צורה לינארית.  
בגרף זה ניתן לראות כי לא קיים פיזור אחיד יחסית של התצפיות ולכן נניח שהנחת הלינאריות לא מתקיימת, כשבודקים לינאריות נרצה שפיזור התצפיות יהיה אחיד מעל ומתחת לציר ה-X.

בבדיקת הנחת **שוויון שונויות** נסתכל על תרשים הפיזור של השגיאות:  
ניתן לראות שאופן פיזור השגיאות על הגרף אינו אחיד, כלומר עבור ערכי x קטנים, השונות נמוכה, אולם עבור ערכי x גדולים השונות גדולה. לכן ניתן לומר שהנחת שוויון השונות לא מתקיימת.

בבדיקת **הנחת הנורמליות** נסתכל על תרשים QQ-PLOT של השגיאות המתוקננות.

תמונה שמכילה טקסט, תרשים, עלילה, צילום מסך

התיאור נוצר באופן אוטומטי

נרצה שבגרף מסוג זה כל התצפיות יתלכדו על הישר Y=X. בגרף זה ניתן לומר שהתצפיות לא מתלכדות על הישר ולכן נוכל להניח שהנחת הנורמליות לא מתקיימת.

* 1. **בדיקה פורמלית של הנחות המודל:**

בכדי לבחון את הנחת **הנורמליות** נבצע שני מבחנים סטטיסטיים בר"מ של 5%, ההשערות הן:

הנחת הנורמליות מתקיימת:

הנחת הנורמליות אינה מתקיימת :

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, גופן, צילום מסך, קו

התיאור נוצר באופן אוטומטי

ניתן לראות שלפי מבחן KS כי  *גם לפי מבחן SW קיבלנו ש ומכאן נוכל לדחות את השערת האפס ונאמר שברמת מובהקות של 5% הנחת הנורמליות לא מתקיימת.*

קיבלנו גם מבחינה ויזואלית וגם מבחינה סטטיסטית כי הנותנים לא מתפלגים נורמאלית.

בכדי לבחון את הנחת **הלינאריות** נבצע את מבחן סטטיסטי CHOW בר"מ של 5%, ההשערות הן:

הנחת הלינאריות מתקיימת:

הנחת הלינאריות אינה מתקיימת :

תמונה שמכילה טקסט, גופן, צילום מסך, קו

התיאור נוצר באופן אוטומטי

ניתן לראות לפי מבחן CHOW כי  *ומכאן נוכל לקבל את השערת האפס ונאמר שברמת מובהקות של 5% הנחת הלינאריות אכן מתקיימת*.

**נספחים:**

2.4.1:

גרפים:

תמונה שמכילה טקסט, צילום מסך, תרשים, קו

התיאור נוצר באופן אוטומטי

עם משתנה אינטרקציה ובלעדיו: weekend

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

עם משתנה אינטראקציה ובלעדיו: holiday

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

עם משתנה אינטרקציה ובלעדיו: לפי עונה

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי