**רגרסיה לינארית - (Linear Regression)**

רגרסיה לינארית היא שיטה סטטיסטית המשמשת למדידת הקשר בין משתנה תלוי לבין משתנה בלתי תלוי אחד או יותר, באמצעות הצגת הקשר כמשוואה לינארית. המטרה היא למצוא את קו ההתאמה הטוב ביותר, כך שהפער בין הערכים החזויים לערכים האמיתיים יהיה מינימלי (למזער את השאריות).

שימושים עיקריים:

* חיזוי (Prediction): שימוש בערכים קיימים כדי לחזות ערכים עתידיים.
* ניתוח מגמות (Trend Analysis): זיהוי דפוסים לאורך זמן.
* זיהוי קשרים (Relationship Identification): הבנת השפעת משתנים בלתי תלויים על משתנה תלוי, למשל, כיצד מחיר משפיע על ביקוש.

במקרה שלנו, השתמשנו ברגרסיה לינארית כדי לחזות מחירי טיסות על בסיס נתונים שונים, כגון מרחק, יום המראה או נחיתה, מספר עצירות, כבודה ועוד.

מאפיינים:

רגרסיה לינארית היא אחת השיטות הפשוטות ביותר ללמידת מכונה ולעיתים משמשת כשלב ראשון לפני שימוש במודלים מורכבים יותר. היא מאפשרת לבחון האם הנתונים מתאימים למודל לינארי ולספק תובנות ראשוניות על הקשרים ביניהם.

**יתרונות וחסרונות של רגרסיה לינארית**

|  |  |
| --- | --- |
| יתרונות | חסרונות |
| * פשוטה להבנה ולפירוש | * לא מתאימה למבנים לא-לינאריים |
| * מהירה ומתאימה לנתונים גדולים | * רגישה לרעשים ומושפעת מערכים חריגים |
| * בעלת יכולת הכללה גבוהה (סיכון נמוך ל-Overfitting) | * פשוטה מדי ולא מתמודדת היטב עם בעיות מורכבות |

מדוע רגרסיה לינארית פחות רגישה ל-Overfitting?

* מודל פשוט מבנית – רגרסיה לינארית תמיד מחפשת קו ישר (או מישור) ולכן אינה יכולה להתאים את עצמה באופן מוגזם לרעש או חריגות בנתונים.
* מספר פרמטרים קטן – מספר המשתנים במודל קטן יחסית, כך שהמודל אינו "זוכר" מידע מיותר ומוגן יותר מפני למידת יתר.
* הגבלה להתאמה לינארית בלבד – מכיוון שהמודל אינו יכול ללמוד יחסים לא לינאריים, הוא מתמקד בזיהוי מגמות כלליות במקום להתאים את עצמו לכל פרט קטן בנתונים.

**שלבי עבודה במחברת:**

שלב 1- בניית מודל ראשוני והערכת ביצועים

בשלב הראשון של יישום רגרסיה לינארית, בחרנו את המשתנים הבלתי תלויים (features) שמשפיעים על מחיר הטיסה, כמו מרחקים, יום ההמראה, כבודה ועוד, והגדרנו את המשתנה התלוי – מחיר הטיסה. לאחר מכן, נרמלנו את הנתונים כדי לוודא שכל המשתנים יהיו בטווח דומה, מאחר שרגרסיה לינארית היא מודל רגיש להבדלים בסקאלה. חילקנו את הנתונים לסט אימון ולסט בדיקה, כך שנוכל להעריך את המודל על נתונים שהוא לא ראה במהלך האימון. אימנו את מודל הרגרסיה הלינארית וחזינו מחירים לטיסות בסט הבדיקה. לאחר מכן, חישבנו מדדי ביצועים כמו MSE, RMSE, MAE ו-R² כדי לבדוק עד כמה התחזיות קרובות לערכים האמיתיים. גילינו שהתוצאות היו חלשות מאוד, עם R² נמוך במיוחד (0.09), מה שמעיד שהמודל אינו מסביר היטב את השונות במחירי הטיסות. לבסוף, הצגנו גרף השוואה בין מחירי התחזיות למחירים בפועל, שחשף פיזור רחב מהקו האידיאלי, דבר שמחזק את המסקנה שהמודל אינו מדויק.

שלב 2- בחירת משתנים מבוססת מתאם ושיפור ראשוני של המודל

בשלב זה, בדקנו את הקורלציה בין כל המשתנים המספריים לבדו למחיר הטיסה, כדי לזהות את המשתנים עם ההשפעה הגדולה ביותר. לאחר ניתוח הנתונים, בחרנו רק את המשתנים עם הקורלציה הגבוהה ביותר כדי לבדוק אם ניתן לשפר את המודל. אימנו מחדש את מודל הרגרסיה הלינארית עם רשימת המשתנים המצומצמת, חישבנו מחדש את מדדי הביצועים והשווינו אותם לתוצאות הראשוניות. אמנם חל שיפור מסוים, עם עלייה ב-R² מ-0.09 ל-0.25 וירידה ב-MSE, אך עדיין מדובר בביצועים נמוכים מאוד. בשלב הבא, ננסה להוסיף משתנים קטגוריים כמשתני דמה (dummy variables) כדי לבדוק אם הם יכולים לשפר את המודל.

שלב 3- הוספת משתנים קטגוריים ושיפור נוסף של המודל

בשלב זה, המטרה הייתה לבדוק אם הוספת משתנים קטגוריים תשפר את ביצועי המודל. לשם כך, המרתנו את המשתנים הקטגוריים "Onward Airlines" ו-"Return Airlines" למשתני דמה (dummy variables) ושילבנו אותם במודל. בנוסף, נורמלנו מחדש את כל המשתנים המספריים כדי להבטיח סקאלה אחידה. לאחר אימון המודל מחדש, חישבנו שוב את מדדי הביצועים וגילינו שחל שיפור נוסף: R² עלה ל-0.28 ו-MSE ירד, אך עדיין מדובר בביצועים לא מספקים. בשלב הבא, נמשיך ללטש את בחירת המשתנים על ידי הסרת משתנים בעלי קורלציה גבוהה עם משתנים אחרים ובדיקת השפעת הוספת משתנים חדשים על דיוק התחזיות.

שלב 4 - בחירת תכונות אופטימלית והסקת מסקנות על הרגרסיה הלינארית

בשלב זה, השתמשנו במטריצת הקורלציה כדי לזהות משתנים בעלי מתאם גבוה עם משתנים אחרים ולהסירם כדי למנוע רב-קו-לינאריות. לאחר מכן, בחנו את השפעת הוספת תכונות שונות על ביצועי המודל, כאשר בכל שלב הוספנו משתנה נוסף ומדדנו מחדש את הביצועים. מצאנו שככל שהוספנו משתנים, R² עלה בהדרגה, וה-MSE ירד, אך השיפור היה מוגבל.

לבסוף, כאשר כללנו את כל 24 המשתנים, קיבלנו את התוצאה הטובה ביותר שניתן להשיג באמצעות רגרסיה לינארית. עם זאת, למרות השיפור, התוצאות עדיין אינן משביעות רצון, מה שמעיד על כך שמערכת היחסים בין הנתונים למחירי הטיסות אינה לינארית. זה מסביר את הביצועים הנמוכים יחסית של המודל, שכן רגרסיה לינארית אינה מסוגלת ללכוד דפוסים מורכבים יותר.

מסקנה זו מצביעה על כך שבשלב הבא יהיה כדאי לנסות מודלים לא-לינאריים, כגון Random Forest, XGBoost, או מודלים מבוססי רשתות נוירונים, כדי לראות האם ניתן לשפר את החיזוי בצורה משמעותית.

שלב 5 - ניתוח שאריות וזיהוי השפעת משתנים על המחיר

בשלב זה, בחנו את דפוסי השאריות (Residuals) של המודל כדי לבדוק האם קיימים דפוסים שיכולים להעיד על בעיות כגון אי-לינאריות. גרף השאריות הראה כי השגיאות אינן מפוזרות באופן אקראי סביב האפס, מה שמחזק את ההשערה שהקשר בין המשתנים למחירי הטיסות אינו לינארי לחלוטין. בנוסף, השווינו בין המחירים בפועל למחירים החזויים, כאשר ניתן לראות שהמודל עדיין מתקשה לתת תחזיות מדויקות במיוחד בטווחים מסוימים.

לאחר מכן, ניתחנו את השפעת המשתנים באמצעות מקדמי הרגרסיה. מצאנו שמשתנים כמו חודש היציאה (Departure Month), חודש החזרה (Return Month), האם מדובר בטיסה בסוף שבוע (Is Weekend Flight), מספר תיקי היד (Carry-ons) ומספר המזוודות (Checked Bags) הם בעלי ההשפעה החזקה ביותר על המחיר. משתנים כמו מספר הקונקשנים בטיסה (Total Connections, Onward Flight Connections) קיבלו מקדמים שליליים, מה שמעיד על כך שככל שמספר הקונקשנים עולה, מחיר הטיסה יורד – ייתכן מכיוון שטיסות עם קונקשנים נוטות להיות זולות יותר מטיסות ישירות.

ניתוח זה מאשר שהתנהגות המחירים אינה לינארית לחלוטין, ומחזק את הצורך בבדיקת מודלים מורכבים יותר, כמו רגרסיה פולינומיאלית או אלגוריתמים לא-לינאריים (Random Forest, XGBoost), כדי לראות האם ניתן לשפר את הדיוק בחיזוי מחירי הטיסות.

**מגבלות הרגרסיה הלינארית בחיזוי מחירי טיסות**

הרגרסיה הלינארית שימשה כמודל בסיסי לחיזוי מחירי טיסות, אך גם לאחר אופטימיזציה, השגתנו R² מקסימלי של ~0.40, כלומר, המודל מסביר רק 40% מהשונות במחירים.

מחירי טיסות מושפעים מגורמים לא-לינאריים ודינמיים, כגון תמחור דינמי של חברות תעופה, השפעות עונתיות, אירועים חיצוניים, וקשרים מורכבים בין משתנים כמו קונקשנים, זמני טיסה וביקוש משתנה ולכן רגרסיה לינארית היא לא מתאימה. המודל אינו מצליח ללכוד את המורכבות של תמחור טיסות, ולכן יש צורך במודלים מתקדמים יותר שיכולים להתמודד עם קשרים לא-לינאריים ולשפר משמעותית את דיוק התחזיות.