**הסבר משימה ד**

במשימה זו התבקשנו לבנות מודל החוזה את הפרש המחירים בין האתרים. צריך להשתמש באלגוריתמים שונים ופרמטרים שונים כדי לנסות לחזות הכי טוב שאפשר.

**הבעיה:** מכיוון שלא כל טיסה באתר אחד זו בדיוק אותו הטיסה באתר השני, וגם אין לכל טיסה משהו שמייחד אותה (כמו ID או מספר טיסה) אז אין לנו באמת איך להשוות את המחיר של טיסה בין האתרים.

**הפתרון:** לכל אתר יש לנו 100 טיסות עבור כל קומבינציה של TTT, LOS, סנאפשוט, מסלול (מקור + יעד) ולכן יש לנו סך הכל 2,700 קומבינציות כאלה. עבור כל פיצ'ר עשינו אגרגציה (לפעמים ממוצע, לפעמים חציון, ולפעמים שכיח – בהתאם למה שנראה הכי הגיוני לפיצ'ר) על כל קומבינציה כזו וכעת הרבה יותר מדויק להשוות בין הטיסות של אתר אחד לבין של השני.

**הערה:** היה ניתן גם לפתור את הבעיה על דרך קלאסטרים אבל מכיוון שכבר השתמשנו בזה במחברת אחרת אז ניסינו גישה אחרת.

לאחר הרצה ראשונית של אלגוריתמים שונים, התוצאות R2 היו כמעת אפסיות בכל האלגו השונים. הסקנו שזה מכיוון שיש לנו רק 70% מ- 2,700 נתונים ולכן אין מספיק מידע בשביל שילמד לחזות כמו שצריך (under-fitting).

לכן חילקנו כל קומבינציה של 100 רשומות לתתי קבוצות של 10 רשומות ולכל קבוצה כזו ביצענו אגרגציה וככה הכפלנו את מספר הרשומות פי 10. המדדים עלו בהתאם.

שקלנו האם כדאי לפצל כל מסלול ולבדוק לפי איך יצרנו בוקס-פלוט וראינו שחוץ מרעש, המסלולים מתפלגים בהפרש מחיר באופן דומה ולכן הוחלט להשאיר אותם יחד (ובנוסף חסכנו הרצות אלגוריתמים על פני 12 חלוקות שונות).

כעת עברנו לשלב **ה – פיצ'ר סלקשן** שבו השתמשנו ב 4 שיטות שונות כדי לפלטר את מפיצ'רים שאינם מועילים ליעילות המודל.

1. Variance Threshold – ככל שהנתונים של פיצ'ר מסווים יותר מגוונים, כך המודל יוכל ללמוד טוב יותר את ההבדלים המשפיעים על המטרה, לכן נבחר להוריד את אילו שבעלי שונות נמוכה.
2. Random Forrest Feature Importance – בעזרת מדד החשיבות המובנה של אלגוריתם הראנדום פורסט, היה נוח להבין אילו פיצ'רים הם אלו שמשפיעים הכי הרבה על המודל ובהתאם סיננו את אלו שלא.
3. Correlation With Target – המדד אולי הכי חשוב. ככל שהקורלציה של משתנה מסוים עם משתנה המטרה גבוה, כך ידע המודל להשתמש בו בשביל לחזות באופן מדויק יותר. בהתאם הורדנו את אלו שבעלי ערך נמוך.
4. K-Best Features – שיטה המריצה חזרות, כאשר בכל חזרה משתמשת ב K פיצ'רים בלבד, ובהתאם ניתן לדעת כמה ואילו מהפיצ'רים שלנו יניבו את התוצאות המדויקות ביותר.

* - היה גם ניתן להשתמש בשיטת RFE המוציאה כל פעם פיצ'ר בודד ובודקת את התוצאה אך הדבר דורש זמני ריצה ארוכים במיוחד ובנוסף הוא בודק את מדדיו בהתאם לאלגוריתם שהוא רץ עליו, בניגוד לשיטה שלנו שבודקת באופן כללי יותר. זה משנה כי אנחנו נריץ מספר אלגוריתמים שונים ולכן חשוב שיהיה כמה שיותר כללי בשלב זה.

לאחר מכן הרצנו את האלגו: XGBoost, Gradiant Boosting and Random Forrest בשל היותם המהירים ביותר וגם מכיוון שהשיגו את התוצאות הטובות ביותר בבדיקה הראשונית.

על כל אחד מהם בוצע gridSearch וקיבלנו ש Gradiant Boosting הוא האלגו הטוב ביותר לשני האתרים, אך עם תוצאות מאוד נמוכות:

Kayak - MSE: 12561.60

Kayak - R²: 0.6589

Kayak - MAE: 82.79

Kayak - RMSE: 112.08

Momondo - MSE: 17752.06

Momondo - R²: 0.5179

Momondo - MAE: 99.73

Momondo - RMSE: 133.24

מסקנות:

✅ התכונות שלנו עשויות להסביר מגמות כלליות בתמחור טיסות, אך לא אסטרטגיות תמחור ייחודיות לכל אתר.

✅ הנחות, עמלות ותזמון עדכונים ספציפיים לכל אתר יוצרים הבדלים בלתי צפויים.

✅ בשל הקושי להעריך את ההפרשים במחירים בין האתרים, נבחן את ההבדלים הכלליים כדי לקבל רמז ברור יותר למגמות.

בחלק השני של המחברת השוונו בין האתרים על פני המון תכונות שונות כנגד המחיר כגון TTT, זמן טיסה כולל, מרחקי של שדה התעופה (המראה/נחיתה/הלוך/חזור) ממרכז העיר ועוד... ולכל אחד יצרנו גרף מתאים ומסקנות.

**סיכום סופי:** מכיוון שמודל החיזוי שלנו אינו מדויק כל כך, בעזרת ההשוואות מהמידע הקיים ניתן להניח ולהסיק המון מסקנות שונות על הבדלי המחירים בין האתרים השונים.