Introduction to Machine Learning (67577) - Hackathon 2023

Hotel Cancellations

Task 1

<u>עיבוד מקדים של הדאטה:</u>

תהליך העבודה:

תחילה חקרנו את הדאטה ואת הפיצ'רים- מה הם מייצגים, מה התחומים עבורם, מי הם המשתנים הקטגוריאליים, וניסינו להגיע לתובנות לגבי מי הפיצ'רים שרלוונטים לפרדיקציה של ביטול הזמנה. ניסינו להבין באיזה אופן עלינו להתמודד עם המשתנים הקטגוריאליים הרבים ואילו פיצ'רים סביר שירעישו את הלמידה ולא יתרמו לה. ניסינו לראות אילו פיצ'רים בעת הסרתם מובילים להגדלה\ הקטנה של השגיאה (על פני האימון) ולהסיק מכך מסקנות לגבי הפיצ'רים שהשארנו ולגבי פיצ'רים חדשים שכדאי לנו ליצור. לדוגמה, ראינו שהוספה של פיצ'ר שמכמת את משך השהייה במלון מתוך תאריכי הצ'ק אין והצ'ק אאוט תרם ללמידה. דוגמה נוספת היא שראינו שמספר המלון שהוזמן העלה באופן ניכר את מספר הפיצ'רים, מה שפגע בתהליך הלמידה. ניסינו לבחון כמה ההורדה של מספר המלון פוגעת בשגיאה וראינו שהיא לא משמעותית כלל, ועל כן שהחסרונות עולים על היתרונות ובחרנו שלא להשתמש בפיצ'ר הנ"ל עבור התחזית.

התמודדות עם פיצ'רים:

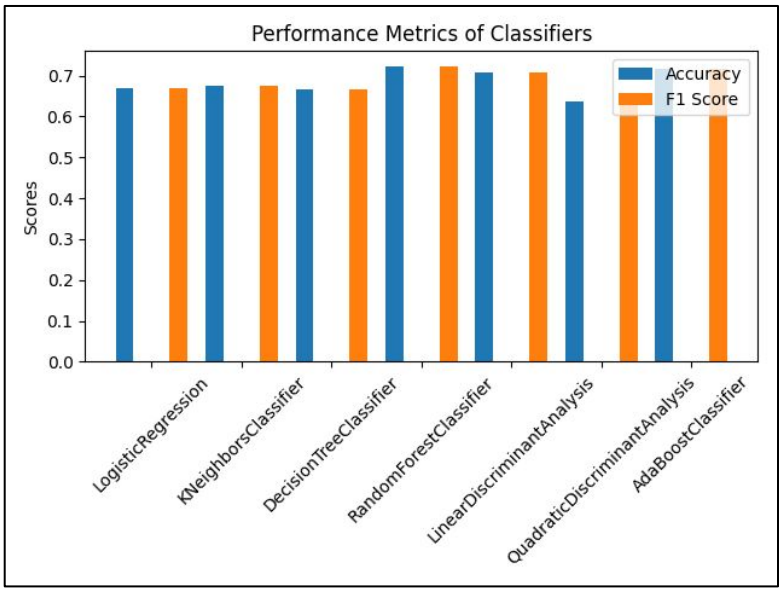
- הפעלנו את הפונקציה get dummies על הפיצ'רים שערכים קטגוריאליים, ולא מוגדר עליהם סדר טבעי.
- כימות של משכי זמן רלוונטיים לפרדיקציה מתוך פיצ'רים שמכילים תאריכים משמעותיים

התמודדות עם ערכים חסרים:

חישבנו את הערכים הממוצעים של סט האימון, ושמרנו אותם יחד עם שמות הפיצ'רים של הסט אימון המעובד. כאשר נתקלים בפיצ'ר חסר הן בסט האימון והן בסט המבחן, החלפנו אותו בערך הממוצע של הפיצ'ר בסט האימון.

<u>בחירת הלומד והיפר-פרמטרים:</u>

בחנו את הביצועים של מספר לומדים לבעיית הקלסיפיקציה עבור הפרדיקציות שהתקבלו לאחר למידת סט האימון, ונראה שהלומד שהשיג את הביצועים הטובים ביותר הוא Random Forest Classifier, ולכן בחרנו בו.



Task 2

<u>עיבוד מקדים של הדאטה:</u>

תהליך העבודה:

כאשר הגענו לבצע את המשימה השנייה, כבר הייתה לנו היכרות יחסית טובה עם הדאטה, בכדי להחליט אילו פיצ'רים רלוונטיים ועשויים לתרום לחיזוי כלשהו. במשימה זאת, ניסינו לחזות את מחיר ההזמנה, לעומת המשימה הקודמת, בה ניסינו לחזות אם יהיה ביטול. על כן ניסינו לעבד את המידע באופן שיותר מותאם לפרדיקציה הזאת. למשל, הוספנו פיצ'ר של החודש בו בוצעה ההזמנה, מכיוון שפעמים רבות ישנם חודשים יותר "נחשקים" (כמו למשל בחופשת הקיץ) בהם המחירים מזנקים, ולכן לפיצ'ר שכזה עשוי להיות ערך לפרדיקציה של מחיר ההזמנה.

ההתמודדות עם ערכים חסרים ועם פיצ'רים במשימה זאת התבצעו באופן דומה למשימה הראשונה.

<u>בחירת הלומד והיפר-פרמטר:</u>

במשימה זאת בחנו לומדים של רגרסיה לינארית. בחרנו בלומד Ridge עם היפר פרמטר חצי. לומד Ridge ממזער את השגיאה תוך ניסיון לבחור את המודל הפשוט ביותר, כך שנמנע מoverfit. בחנו את הביצועים של מודל Lasso, שגם כן ממזער את השגיאה תוך ניסיון לבחור את המודל הפשוט ביותר, אך ראינו שביצועיו פחותים ביחס ל-Ridge.

את ההיפר פרמטר עבור לומד ה-Ridge בחרנו תוך למידה של cross validation ובחינה של ערכים שונים, ולבסוף בחירה של הערך שהביא לשגיאה הנמוכה ביותר.