Project.pdf

מאפיינים מאתגרים של הדאטאסט:

הדאטאסט מכיל הרבה features, במיוחד אם ממירים את כל המשתנים קטגוריאליים למשתנים דינאמים.

לחלק מהמשתנים קורלציה נמוכה עם cancellation, ויש כפילויות (למשל מדינת מוצא, ושפה או hotel_chain ו-hotel_brand).

data cleaning:

- המרנו את cancellation_datetime למשתנה בוליאני (1 אם ההזמנה בוטלה).
- המרנו את cancellation_policy_code לעמודות שמתארות את המידע בצורה מספרית.
- השמטנו משתנים עם קורלציה נמוכה עם y, למשל hotel_live_date, no_of_room, no_of_extra_bed.
- יצרנו משתנים חדשים עם קורלציה גבוהה ל-y: משך החופשה בימים (ההפרש בין תאריך ה-check-in ל-check-out), ומס' הימים בין יום ההזמנה ליום ה-check-in.

booked_days_before	0.309536
duration_days	0.137351

- במקום ליצור משתני דאמי לכל הקטגוריות, לאחר ששמנו לב כי חלק מהקטגוריות בקורלציה נמוכה עם y, יצרנו משתני דאמי רק לערכים שהיתה להם קורלציה גבוהה יחסית עם y. למשל, עבור origin_country_code רק לכמה מהמדינות עם הקורלציה הגבוהה ביותר

origin_country_code_KR	0.095353
origin_country_code_MY	-0.086316
origin_country_code_TH	-0.084208
origin_country_code_TW	0.080084

בדומה מיפינו לקוחות בעייתיים ויצרנו להם דאמי, ומיפינו מדינות שבהן נוטים לבטל (או לחלופין לא לבטל).

שמנו לב שיש מאפיינים שונים להזמנות עם סיכוי גבוה יותר להיות מבוטלות לעומת הזמנות שאינן מבוטלות, למשל יש יותר סיכוי לביטול כאשר הלקוח logged in, ויש סיכוי נמוך לביטול כאשר אופן התשלום הוא pay now לעומת תשלום דחוי. לכן גם בדקנו שימוש במודלים כגון KNN. (*ראו גרפים 1-2 המצורפים בנספח)

על מנת לנסות ולשפר את הנתונים הרצנו את אלגוריתם PCA על מנת להוריד מימד:
כדי לבחור את K חישבנו את הע"ע של מטריצת ה-sample covaraince ובחרנו את כל הע"ע ששונים מ0.
לאחר מכן יצרנו את גרף 3 (* ראו נספח) על מנת לבחור את K.

ניתן לראות כי באזור הערך 4 הגרף משתטח- כלומר הע"ע מתקרבים לאפס, לכן בחרנו בערך זה כמימד עבור הPCA. עם זאת, הנתונים שקיבלנו עם הרצת המודלים על הדאטה במימד החדש היו פחות טובים ובחרנו לא להשתמש בו. ייתכן כי הדבר נובע מכך שהדאטה עדיין רועש וPCA עלול להעצים זאת.

סה״ משום שיש הרבה מאפיינים למידע בעלי קורולציה נמוכה שלא מייצגים טוב האם עסקה תבוטל אנחנו מצפים שה generalization error יהיה יחסית גבוה אך נמוך מהמודל הבסיסי בו אנחנו מנבאים שכל עסקה לא תבוטל.

התחלנו בבדיקת מודל בסיסי ביותר שחוזה 0 לכל הדגימות על מנת לקבל הערכה כללית לשגיאה ממנה נרצה להשתפר. קיבלנו f1=0.516

מודלים שניסינו:

לצורך מודל הקלסיפיקציה, השתמשנו ב- Logistic Regression, KNN, DecisionTree.

AdaBoostClassifier עם עץ החלטה :

לצורך קביעת מספר הלומדים וגובה עץ ההחלטה השתמשנו בCross Validation ובחרנו בפרמטרים בעלי הscore הטוב ביותר:

```
{'base_estimator': DecisionTreeClassifier(max_depth=1), 'n_estimators': 200}
```

התוצאות שקיבלנו במודל זה:

F1 macro score: 0.732, accuracy: 79.55%

עם KNeighborsClassifier קיבלנו

1 neighbors - Train accuracy: 0.99, Test accuracy: 0.73, F1 macro score: 0.66
5 neighbors - Train accuracy: 0.83, Test accuracy: 0.76, F1 macro score: 0.679
25 neighbors - Train accuracy: 0.79, Test accuracy: 0.78, F1 macro score: 0.696

עם LogisticRegression קיבלנו

Accuracy: 0.787, F1 macro score: 0.691

עם DecisionTreeClassifier קיבלנו

Max depth: 1,Accuracy: 0.76, F1 macro score: 0.69
Max depth: 5, Accuracy: 0.77, F1 macro score: 0.699
Max depth: 10, Accuracy: 0.795, F1 macro score: 0.732
Max depth: 20,Accuracy: 0.76, F1 macro score: 0.691
Max depth: None,Accuracy: 0.74, F1 macro score: 0.674

לבסוף בחרנו במודל HistGradientBoostingClassifier שבו קיבלנו את התוצאה הטובה ביותר עם:

```
Average cross validation score: 0.805

Test accuracy: 0.808

F1 score: 0.746
```

הרחבה לגבי Task2:

התחלנו בהרצת מודל בסיסי של mean price על מנת שנוכל לקבל הערכה בסיסית לשגיאה הערכים שקיבלנו סביב RMSE= 370.
אחרי הרצות של מספר מודלים הגענו ל RMSE=220 באמצעות מודל רגרסיה HistGradientBoostingRegressor עם מודל בסיס של decision tree regressor, שנתן לנו את התוצאה הטובה ביותר משאר מודלי הרגרסייה הממומשים בקוד.
האלגוריתם הוא אלגוריתם boosting היוצר היסטוגרמה לכל פיצ'ר ומשתמש בה כדי למצוא את הערכים הטובים ביותר לחלוקת הdata. המודל מקבל מספר רב יחסית של היפר פרמטרים. השתמשנו בCV על חלקם כדי לבחור את המודל הטוב ביותר. הערכים שנבחרו:

```
{'learning_rate': 0.14402245736326458, 'max_depth': 3, 'max_iter': 112}
```

נוסף על כך, בנסיון לשפר את התוצאות ביצענו דאטה סקיילינג לאחר הpre-process באמצעות מודל standard_scaler. המודל מבצע סטנדרטיזציה של הפיצ'רים ע"י מיצוע סביב 0 ושונות 1. בצורה זו דואגים שכל הנתונים יהיו בטווח נורמלי. לצערנו, הדבר לא הביא לשיפור בביצועים.

לבסוף הרצנו את המודל ההתחלתי וחזינו איזה הזמנה תבוטל או לא, אם תבוטל אז הכנסנו את ערך ההזמנה לעמודה אחרת -1 כנדרש והחזרנו את הטבלה הנדרשת במשימה.

נספח גרפים ל-project.pdf: