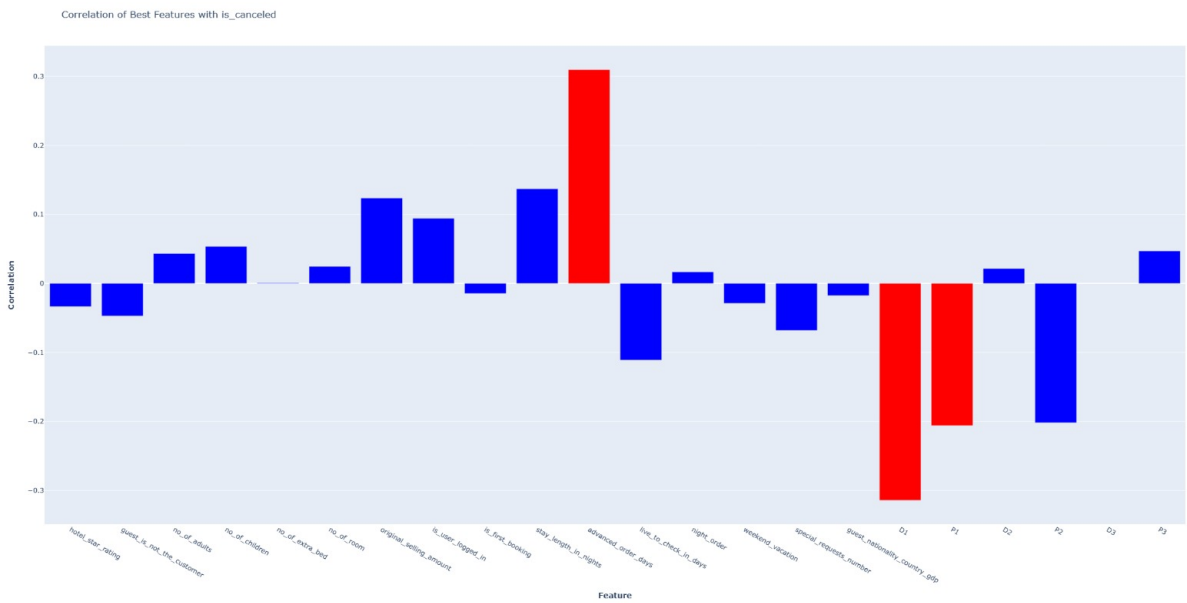
האקתון – משימה 3 - Churn prediction Model

בעיבוד המקדים שבצענו לדאטא זיהינו 3 פיצ׳רים בעלי השפעה יחסית רבה על הפרדיקציה. שלושתם אינם פיצ׳רים שהיו בדאטא במקור ובנינו אותם על בסיס הנחה שישפיעו על קבלת ההחלטות של הלקוח:

א. D1 – פיצ׳ר זה מייצג כמה ימים מראש ניתן לבטל את ההזמנה ללא תשלום.
ב. P1 – פיצ׳ר זה מייצג את מחיר הביטול במידה ומבטלים את ההזמנה בטווח שקטן מ-D1.
ג. Advanced_order_days – פיצ׳ר זה מייצג כמה זמן מראש הוזמן המלון, כלומר מרווח הזמנים בין תאריך הבוקינג לתאריך הצ׳ק-אין.

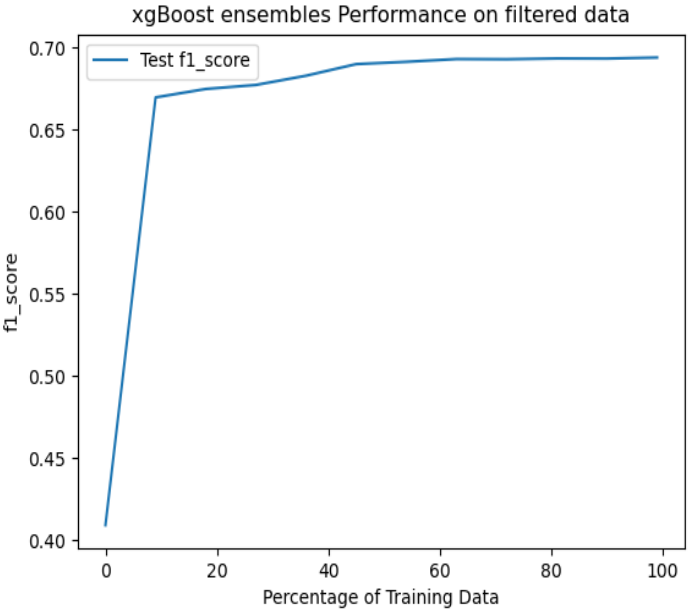
בנינו את שני הפיצ׳רים הראשונים מתוך הבנה כי למדיניות הביטולים של המלון תהיה השפעה על ההחלטה של הלקוח האם לבטל את הזמנתו או לא. פיצלנו את הפיצ׳ר cancellation_policy_code למספר פיצ׳רים שונים המייצגים חלקים חשובים בראייתנו במדיניות הביטולים. מתוך אלו, זיהינו את שני הפיצ׳רים (D1, P1) כמשפיעים ביותר מביניהם.

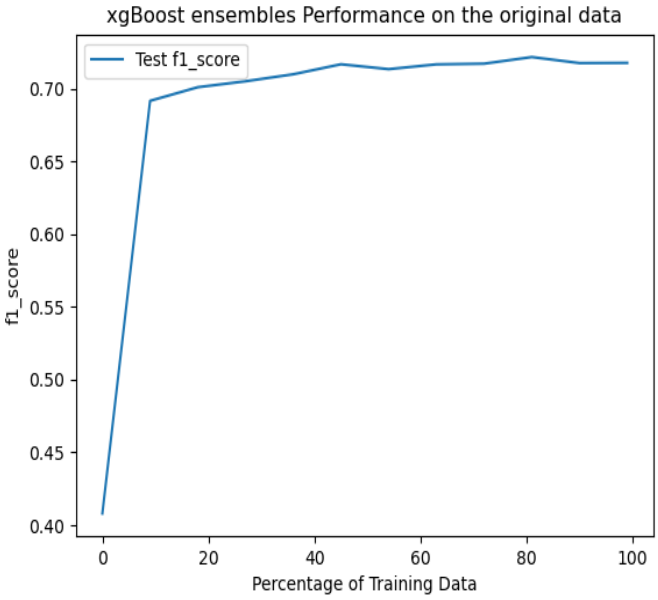
נתבונן בגרף הבא:

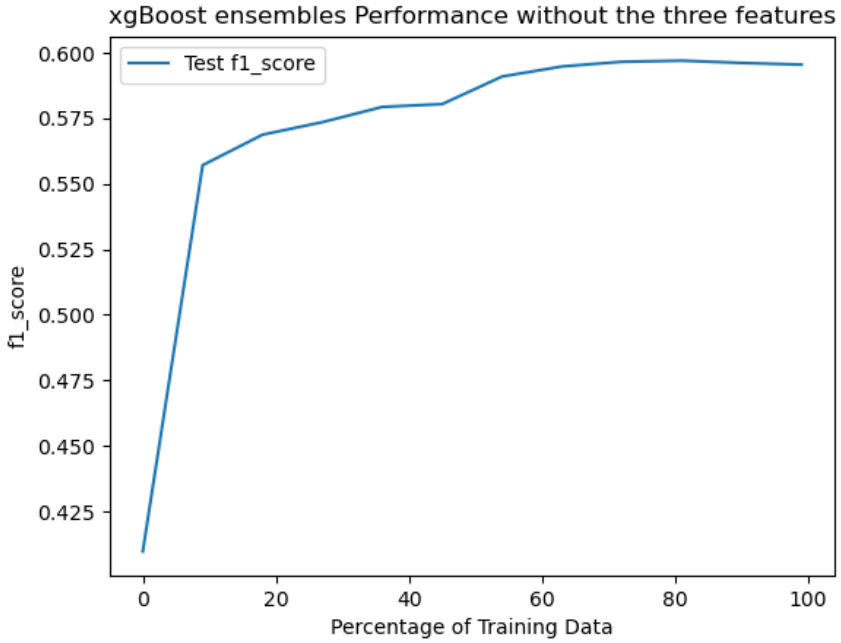


גרף זה מייצג את הקורלציה בין כל אחד מהפיצ׳רים הנמדדים בסט האימון לאחר העיבוד, לבין הפיצ׳ר "is_canceled" – פיצ׳ר בינארי שבנינו על מנת לייצג הזמנות שבוטלו. באדום מסומנים הפיצ׳רים הדומיננטיים לגביהם פירטנו. נשים לב כי הפיצ׳ר Advanced_order_days הוא בעל הקורלציה החיובית הגבוהה ביותר עם הפיצ׳ר "is_canceled" – כלומר יש קשר חיובי חזק (ביחס לשאר הפיצ׳רים) בין ביטול ההזמנה לכמה זמן הוזמנה מראש. בנוסף, יש קשר שלילי חזק (ביחס לשאר הפיצ׳רים) עבור D1 ועבור P1, כלומר ככל שמספר הימים לביטול לפני תשלום נמוך יותר וכן ככל שהמחיר לתשלום בעת ביטול נמוך יותר, כך פחות סביר שהלקוח יבטל את הזמנתו.

בהתאם לכך, נתמקד כעת בשלושת הפיצ׳רים האלו. נתבונן בגרפים הבאים:







בגרף השמאלי העליון הורדנו את כל הפיצ׳רים מסט האימון למעט שלושת הפיצ׳רים בהם אנו דנים. לאחר מכן, אימנו את מודל ה- xgBoost על גדלים שונים של הדאטא המעובד, ביצענו פרדיקציה על סט המבחן וחישבנו f1_score. נשים לב כי f1_score על דאטא זה מאוד דומה ל-f1_score שנשיג על הדאטא המקורי (מופיע בגרף השמאלי) - בשני המקרים אנו מזהים התייצבות של f1_score על ערך יחסית קרוב ל0.7, במגמה די דומה. לכן, נוכל להסיק כי שלושת הפיצ׳רים מייצגים את הקשר בין הדאטא לסיווג באופן משמעותי. לבסוף, בגרף התחתון הסרנו את שלושת הפיצ׳רים מהדאטא המקורי וחזרנו על התהליך. ניתן לשים לב שקיבלנו ערך f1_score נמוך ביחס לניסויים הקודמים, ולכן נוכל להסיק כי הסרת הפיצ׳רים פגעה ביכולת של המודל לבצע פרדיקציה.