**פרויקט כריית מידע וייצוג מידע**

**פרויקט סוף**

תאריך: 2024\7\21

מגישים: יקיר מרמלשטיין ושהם קראוס

ת.ז: 325694081, 207231937

ראשי פרקים

Case problem and mission

Data information

Data cleaning

Data statistics

Data correlation and visualization

Data manipulation

Data preprocessing

Data modeling and hyperparameters

Data training and model evaluation

Test Data preprocessing and prediction

**Case problem and mission**

תיאור הבעיה:

בתי מלון לעתים קרובות מתמודדים עם אתגרים בניהול הזמנות בשל ביטולים בלתי צפויים. ביטולים אלה עלולים להוביל לאובדן הכנסות והקצאת משאבים לא יעילה. חיזוי מדויק של ביטולי הזמנות הוא חיוני לבתי מלון כדי למטב את שיעורי התפוסה ולהגדיל את הרווחיות.

משימתנו:

בפרויקט זה, עלינו לנתח מערך נתונים של הזמנות מלון כדי לפתח מודל שיכול לחזות אילו הזמנות סביר שיבוטלו. זה יאפשר לבתי מלון ליישם אסטרטגיות הזמנה יעילות יותר ולשפר את היעילות התפעולית הכוללת.

תחילה נחקור וננתח את מערך הנתונים לאימון -'hotels\_train' באמצעות שיטות סטטיסטיות וכלי ויזואליזציה, ונבצע עיבוד מקדים.

לאחר מכן, נאמן מודל סיווג על מערך הנתונים המעובד ונבצע תחזית עבור מערך הנתונים לבדיקה המצורף -'hotels\_test'.

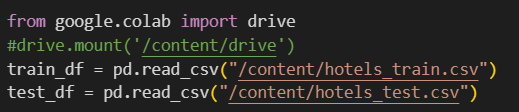
**Dafa information**

עבדנו בסביבת העבודה -Colab

תחילה ייבאנו את הספריות המתאימות הנוגעות לעיבוד נתונים,הצגת נתונים והפעלת המודלים השונים

ביניהן numpy ,pandas,matplotlib,sklearn,torch

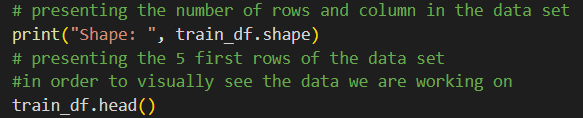
לאחר מכן ייבאנו את הקבצי הדאטה למסמך העבודה

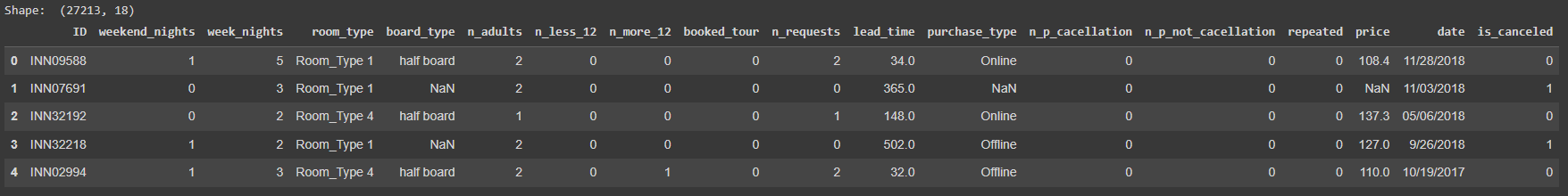


כעת על מנת לעבוד עם הדאטה יש צורך להבין כיצד הדאטה מאורגן, לשם כך הצגנו את 5 הרשומות הראשונות בקובץ האימון באמצעות פקודת head.

בנוסף הדפסנו את מספר השורות -הרשומות בדאטה האימון ואת מספר העמודות- הפרמטרים ממנו הוא מורכב

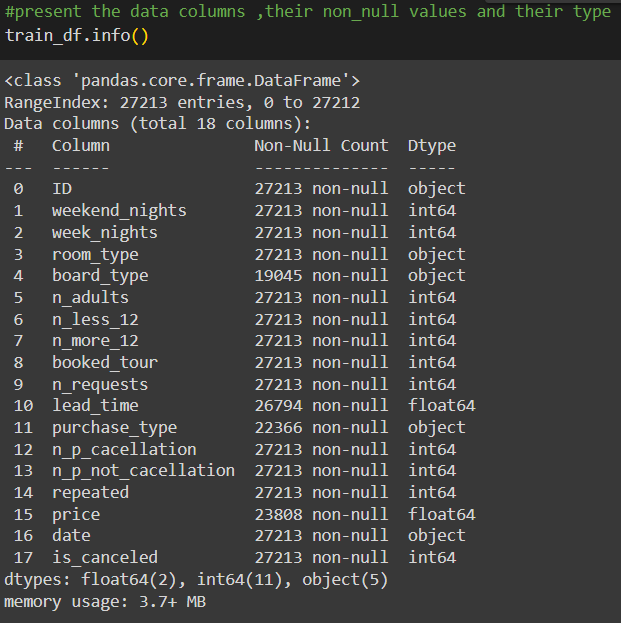
קיבלנו מאגר נתונים המכיל כ18 עמודות וכ- 27213 שורות.





כמו כן השתמשנו בפקודת Info בכדי להציג בצורה מסודרת את עמודות הטבלה כדי שנוכל לדעת את ה-Data type עבור כל עמודה בצורה נוחה .(בנוסף ניתן לראות כי חלק מהעמודות לא שלמות , נבחן זאת בהמשך באמצעות פונקציית isnull).

ניתן לראות כי סוגי העמודות הינם מסוג : float,int ,object .



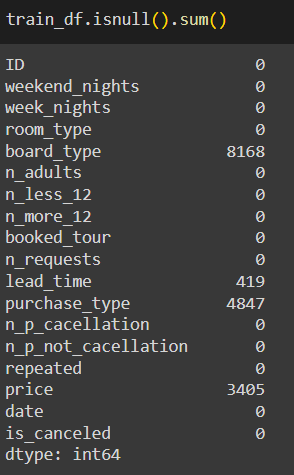
**Data cleaning**

1.הורדת העמודות הלא נחוצות:

* באמצעות פקודת drop הורדנו את העמודה של ה -ID



2. טיפול בחוסר התאמות:

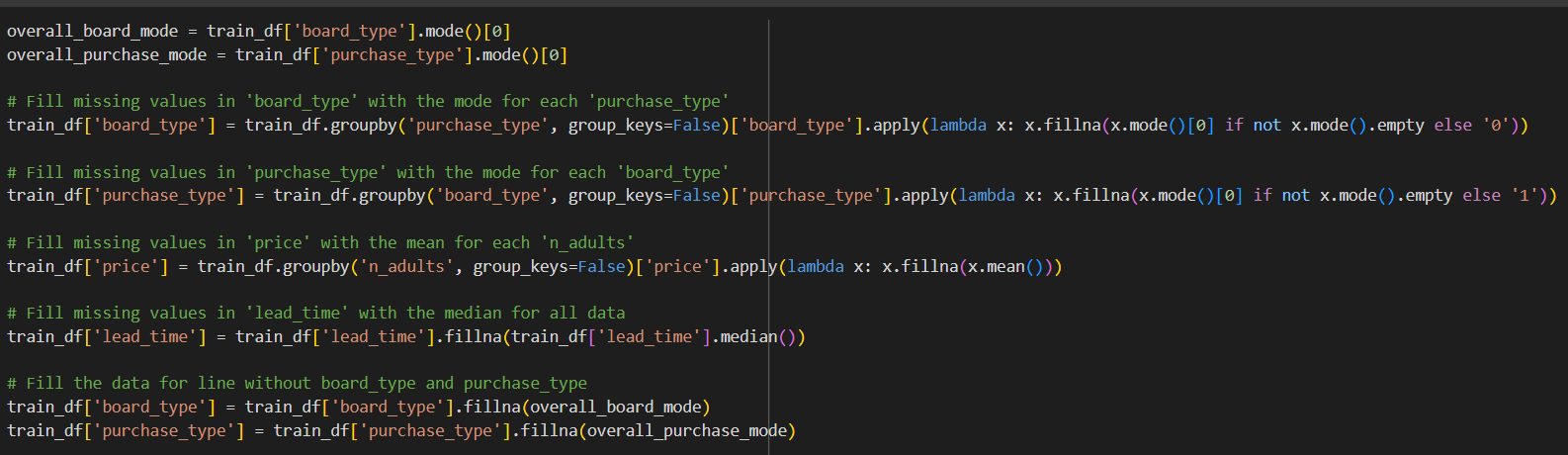
* זיהינו את התאריך '2018-2-29' תאריך שאינו קיים בשנה הקלאנדרית 2018 כיוון שמדובר מספר נמוך של רשומות מתחת ל12 , בחרנו להסיר אותן מקובץ הtrain



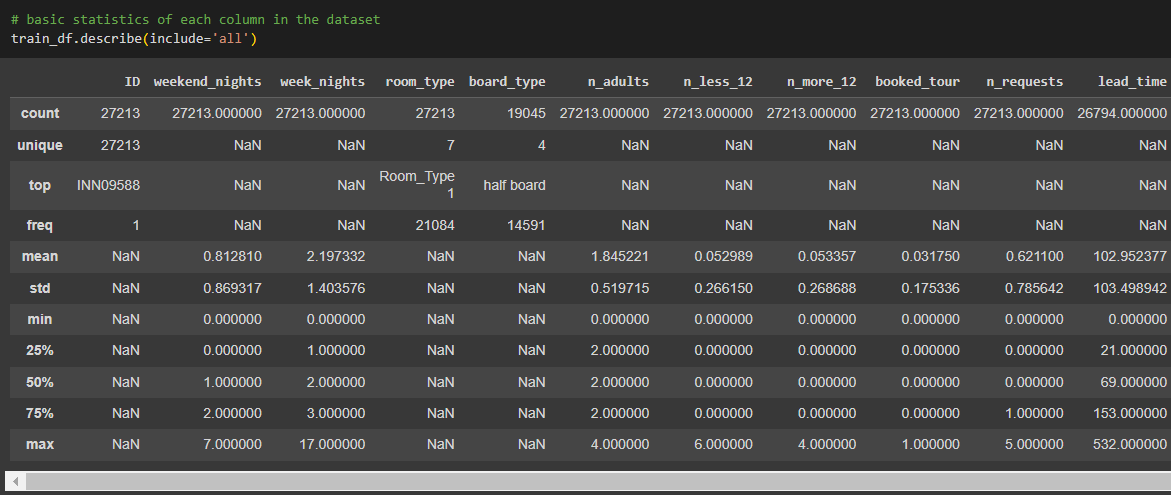
נבחן את כמות הערכים החסרים בדאטה האימון ביחס לעמודות באמצעות isnull

3. טיפול בערכים חסרים:

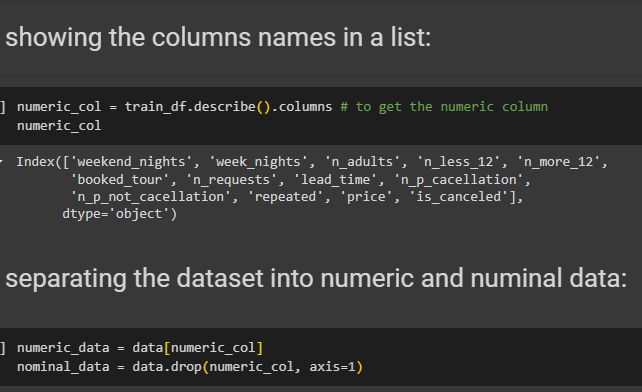
* בשלב הצגת הנתונים זיהינו חוסרים משמעותיים ב4 קטגוריות ( עמודות)
* 'Board\_type’ - שחסרים לנו כ8168 ערכים השלמנו את הערכים באמצעות
* נשלים את הערכים החסרים ב-board\_type באמצעות הערך הנפוץ ביותר בקבוצת purchase\_type המתאימה.
* 'lead\_time'- שחסרים לנו כ419 ערכים
* נשלים את הערכים החסרים ב-lead\_time עם הערך הממוצע הכללי של העמודה lead\_time.
* 'purchase\_type’ - שחסרים לנו כ4847 ערכים
* נשלים את הערכים החסרים ב-purchase\_type באמצעות הערך הנפוץ ביותר בקבוצת board\_type המתאימה.
* כאשר חסר הערך גם בעמודת ה-purchase\_type וגם בעמודת ה-board\_type נשלים כל ערך חסר שנותר בעמודות board\_type ו-purchase\_type עם הערכים הנפוצים ביותר שנשמרו קודם לכן.
* 'price’ - שחסרים לנו כ3405 ערכים
* נשלים את הערכים החסרים ב-price באמצעות הערך הממוצע בקבוצת n\_adults המתאימה.



**Data statistics**



הצגת טבלת סטטיסטיקה:

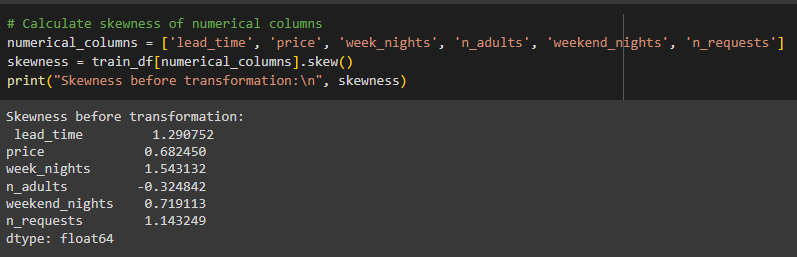


הפרדנו את עמודות הדאטה

ל2 קטגוריות מספרי וקטגורי :

צידדויות

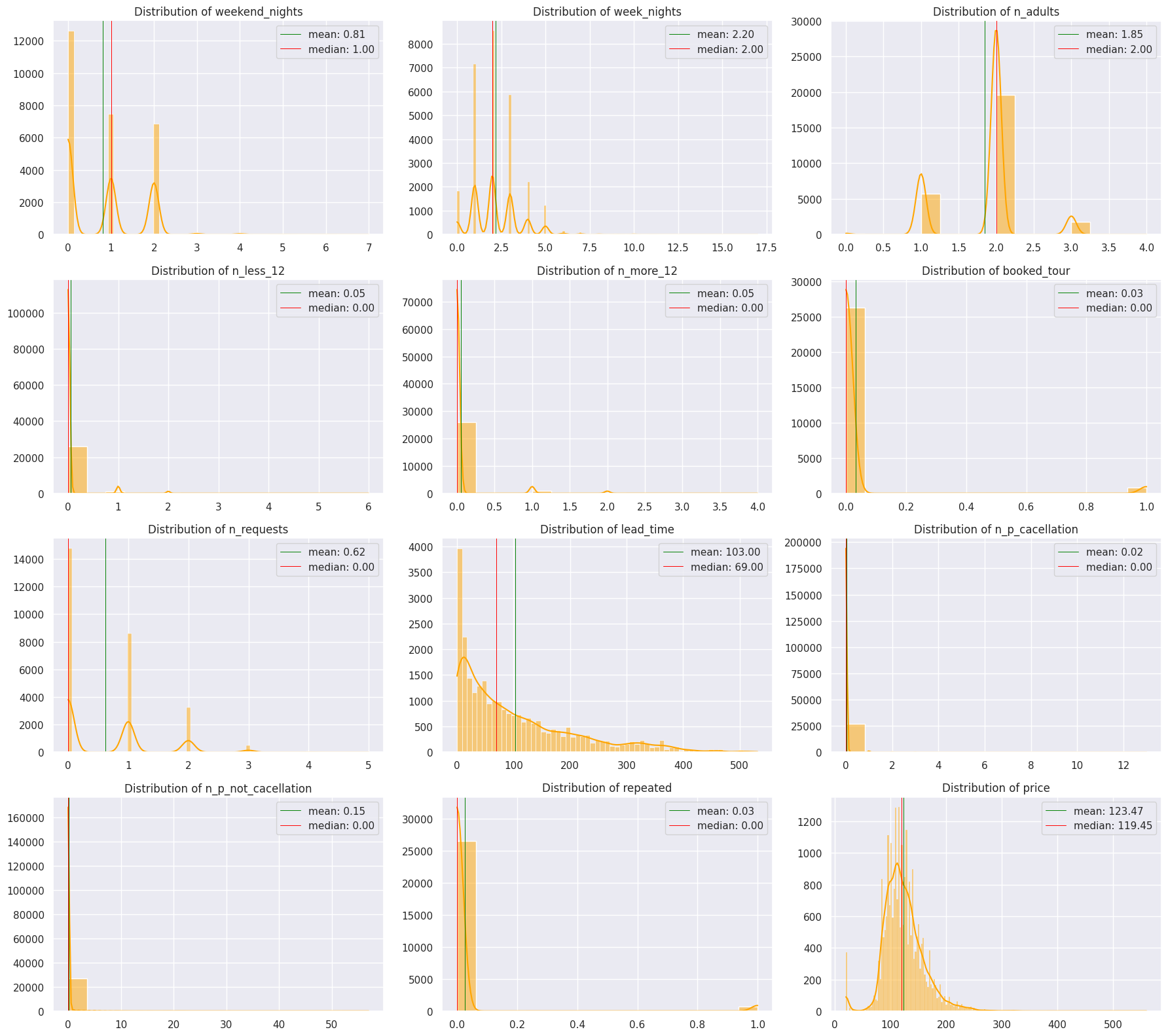
עבור הקטגוריות המספריות חישבנו את ה-Skewness



לצורך שלב הצגת הנתונים הסטטיסטים ביצענו ויזואליזציה שמשלבת בתוכה בצורה ויזואלית נראית לעין את גרפי ההתפלגות השונים עבור כל אחת העמודות בדאטה, בשילוב הצגת הממוצע והחציון עבור הדאטה באמצעות קווים אנכיים. כמו כן ניתן לראות דרך ההצגה את ההבדל בין הדאטה המספרי לדאטה הקטגורי .

*כאשר בציר ה-Y מצויין מספר המופעים של אותו ערך ברשומות הדאטה, ובציר -X הערך בעמודה עבור אותו מופע.*

*התפלגויות משולבת ממוצע וחציון*



(בנוסף ביצענו אחד מובחן יותר ללא סימוני חציון וממוצע.)

**Data correlation vizualization**

עבור כלל העמודות בדאטה ביצענו הצגה ראשונית של קורלציה באמצעות ספריית Seaborn

בHeat map correlation matrix

מטריצת קורלציה של Seaborn heatmap מספקת תובנות חשובות לגבי הקשרים בין מאפיינים במערך נתונים:

זיהוי קשרים חזקים: עוזרת בזיהוי מאפיינים בעלי קורלציה גבוהה שיכולים להעיד על כפילות מידע.

זיהוי מולטיקולינאריות: מזהה מאפיינים שמראים קורלציה גבוהה ביניהם, מה שיכול להוות בעיה עבור מודלים מסוימים.

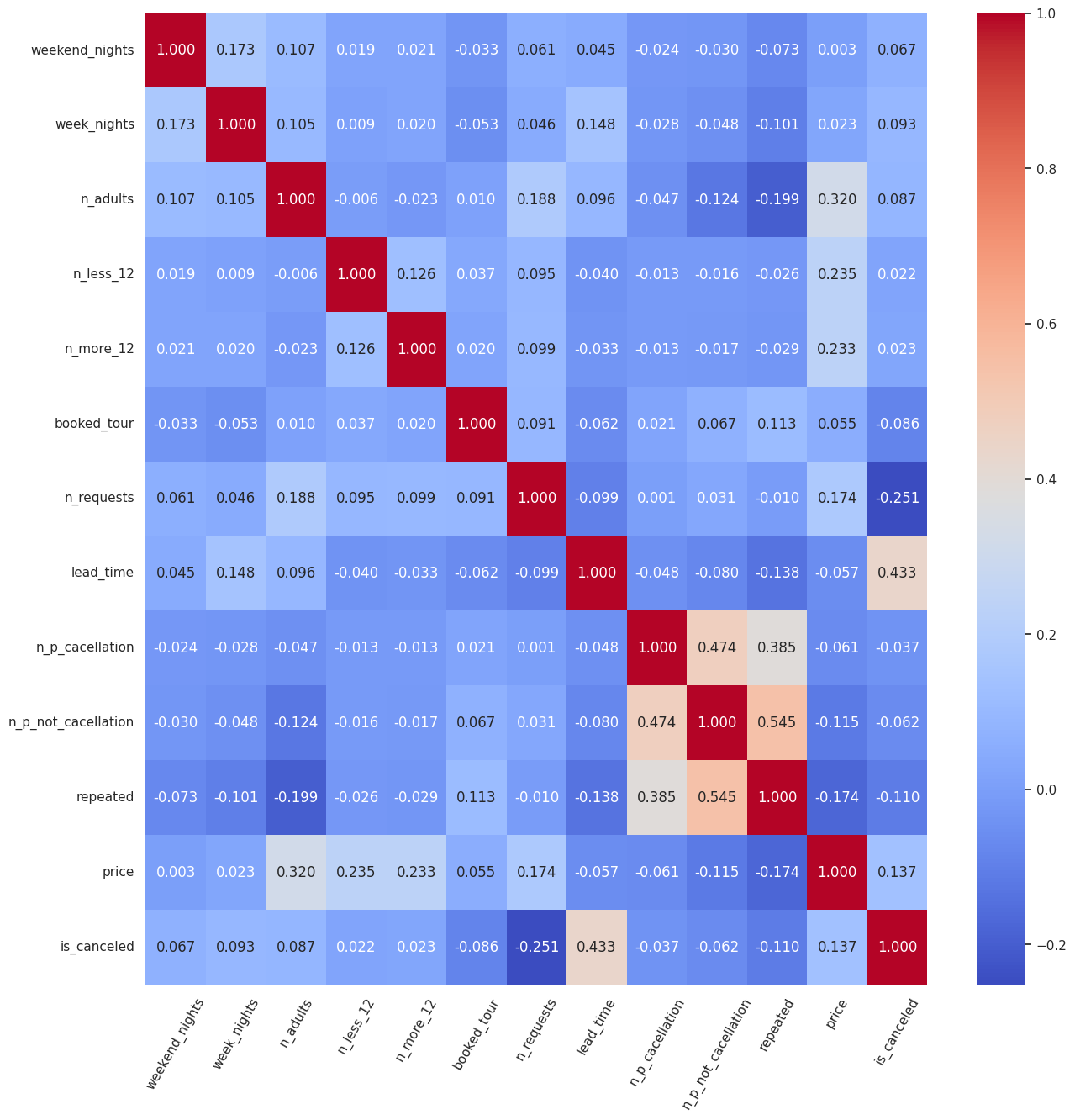
בחירת מאפיינים: מסייעת להחליט אילו מאפיינים לשמור או להסיר בהתבסס על הקורלציה שלהם עם המשתנה המוסבר.

הנדסת מאפיינים: מנחה ביצירת מאפיינים חדשים או בהסרת מאפיינים מיותרים.

הבנת מבנה הנתונים: מספקת תחושה ויזואלית של המבנה הכללי והתפלגות הקורלציות.

ניקוי נתונים:עוזרת לזהות שגיאות נתונים או חריגות ומדגישה את הצורך בהשלמת ערכים חסרים.

תובנות מודל: מספקת תובנות לגבי חשיבות המאפיינים והקשרים שהמודל מנצל.



*ככל שהערכים בטבלה יותר קרובים למינוס 1 או ל 1 הם יותר מתואמים.*

*ניתן לראות כי כל ערך מתואם עם עצמו ב100% ( מן הסתם) ומוצג בצבע אדום בוהק תחת הערך -1.*

ניתן לראות שיש קשר בין החישוב של האם הלקוח הוא לקוח חוזר להאם בערך היו לאותו הלקוח הזמנות והוא ביט אותן או שמה לא ביטל . ( כיוון שאם הוא הזמין בעבר Repeated אזי בהכרח הוא ביטל או לא ביטל הזמנה קודמת)

כיוון שאין מספר עמודות מרובה ונראה חוץ מהמקרה המצויין למעלה שאין כפילוית בדאטה

וכן כיוון שחישוב ומציאת הערכים הריקים והשלמתם עבור העמודות בעלות הערכים החסרים נעשה בדרך אחרת ויותר פשוטה.

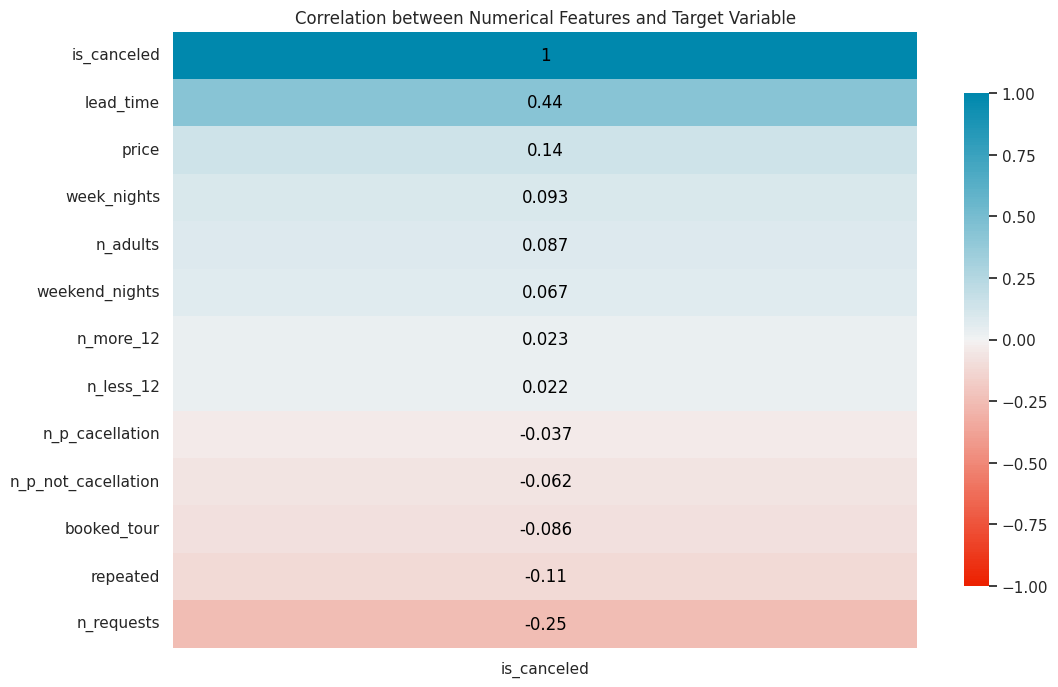
ערכים חשובים שבדקנו במקרה שלנו באמצעות ההצגה הטבלאית הזו על מנת להבין טוב יותר את הטבלה

הינם הקורצליות בין העמודות השונות לערך המטרה בשורה האחרונה

שם ניתן לראות את חשיבותבין היתר של ה-lead time ,number of requests ,price כעמודות חשובות ביחס לפרדיקציה שאנו נדרשים לה האם הלקוח יבטל או לא יבטל את ההזמנה שלו.

כעמודות שערכן רחוק בטבלה מערך ה-0.

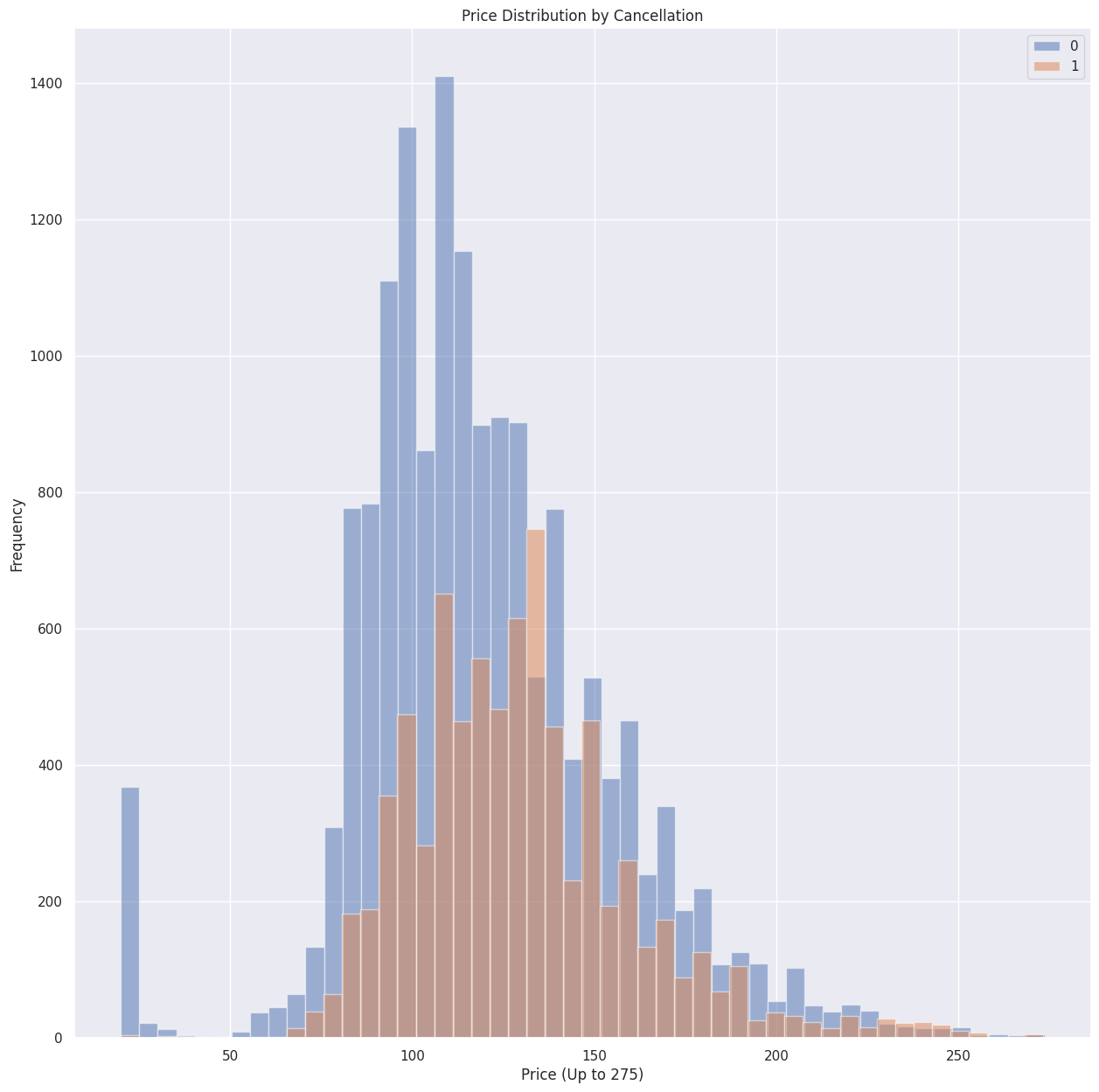
*כיוון שכך יחדנו לכך ויזואלזציה נוספת:*



*כיוון שמדובר בערכים מספריים רציפים בחרנו להציג את Lead time ואת Price גם בהיסטוגרמה רציפה*

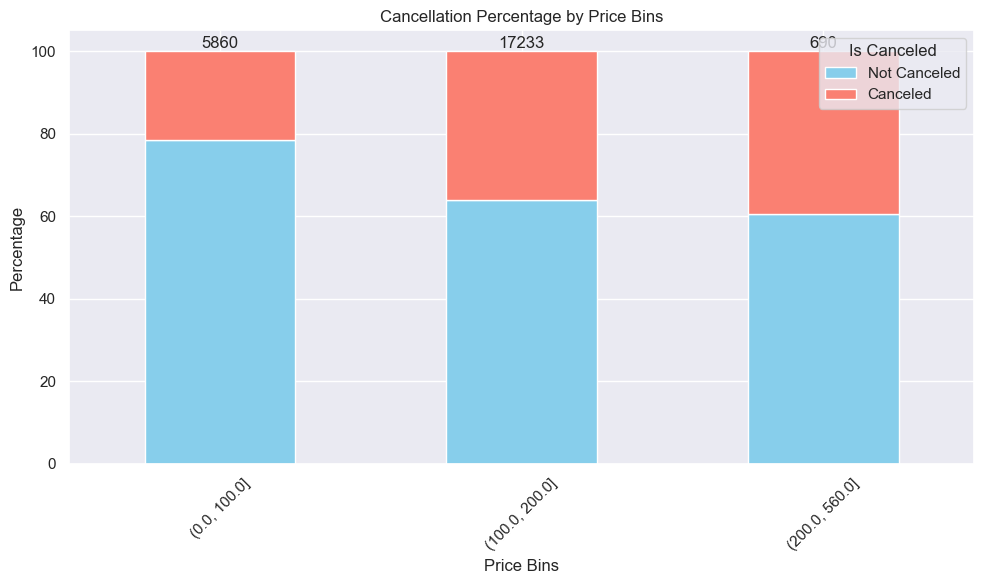
*המחולקת למקטעים של 10 , המציגה באדום את כמות הערכים בטווח הנתון שעבורם בוטלה ההזמנה(1).*

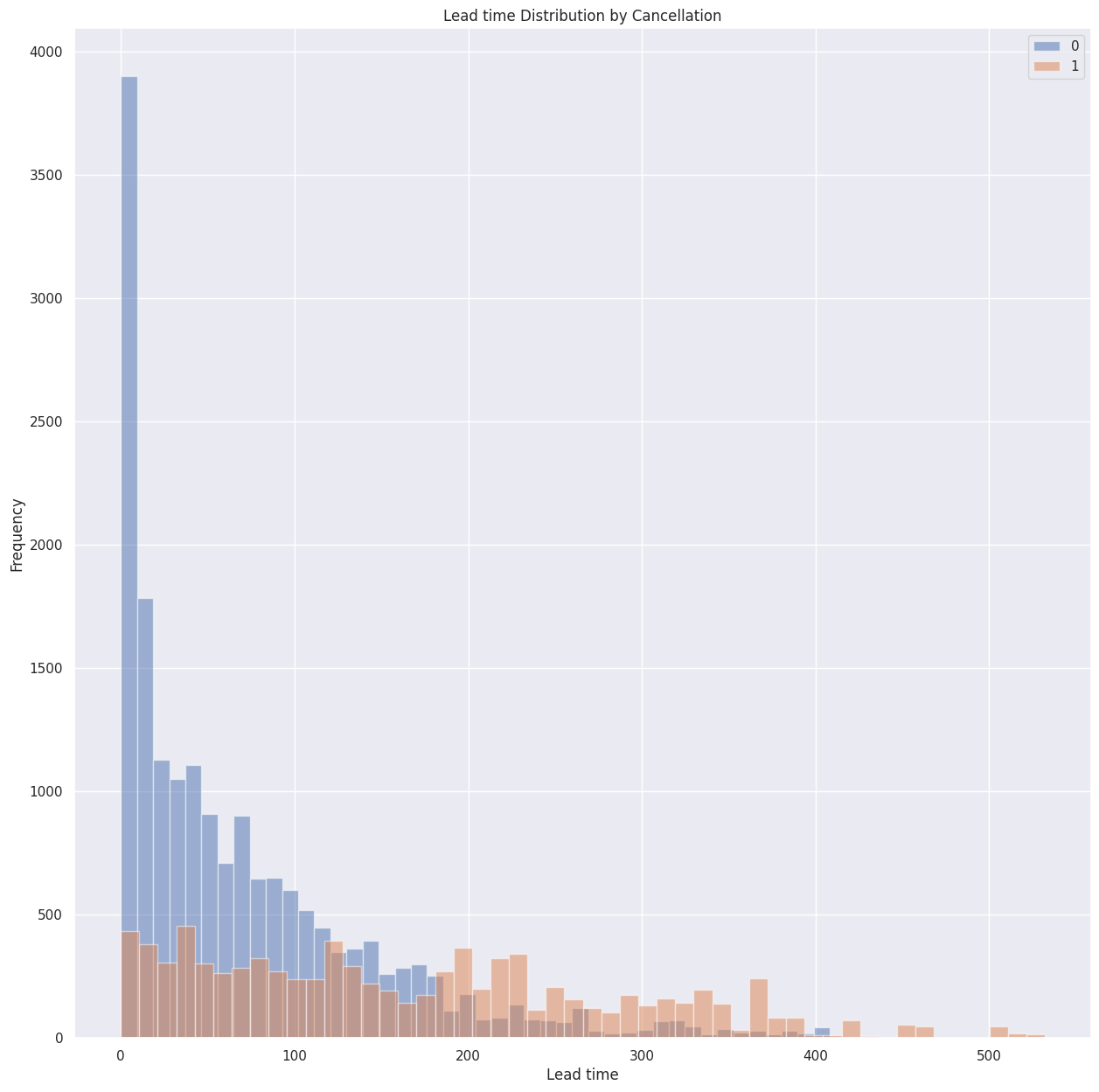
*לעומת ההיסטוגרמות בצבע הכחול שמציינת כמו ערכים בטווח עבורם ההזמנה לא בוטלה (0).*



*ניתן לראות שישנה הטייה מסוימת עבור ביטולים כתלות במחיר, מחיר גבוה גורר יחס גבוה יותר של ביטולים*

*נראה זאת גם באמצעות הגרף הבא. בוא הצגנו זאת ב3 גרפי עמודה המחולקים לBINS - טווחי מחיר*

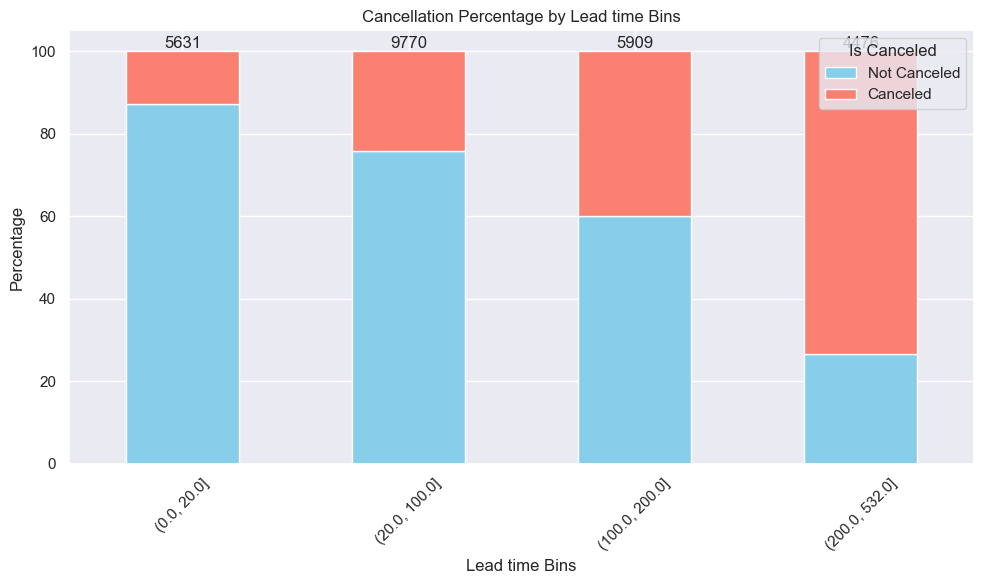




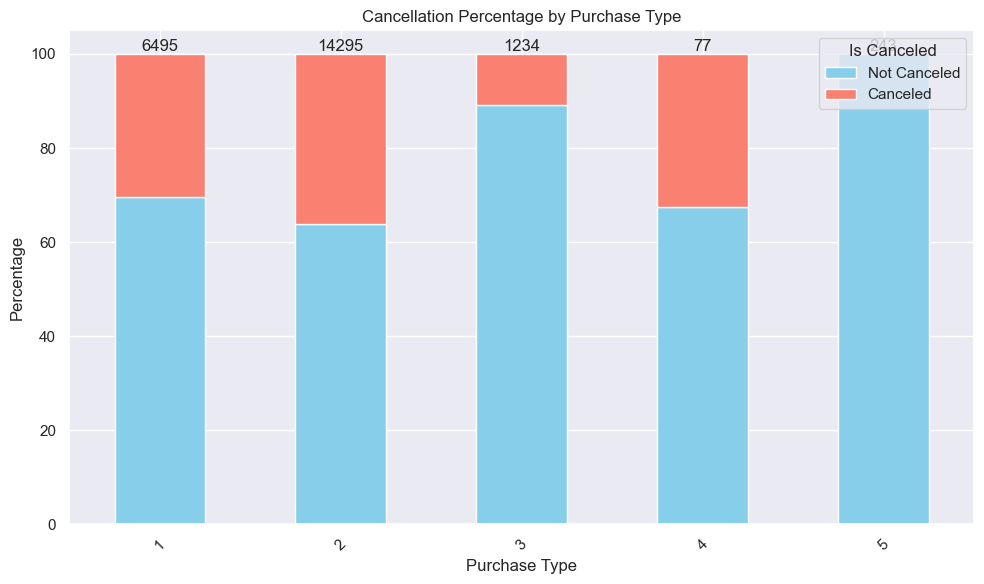
*בנוסף ניתן לראות כאן את מה שראינו בטבלת הקורלציה - על הקשר החזק בין הזמן מביצוע ההזמנה לתאריך שנשמר.*

*כך שעבור זמני המתנה קצרים אחוז הביטולים קטן יחסית בעוד ככל שזמן ההמתנה גדל באופן כללי,והיחס אפילו מתהפך כך שכמות הביטולים גדולה מכמות אי- הביטולים !( מאיזור 200 ימים והלאה).*

*גם כאן ההצגה ב bins עבור הLead time , עוזרת לראות במפורש את שינוי המגמה ביחס הביטולים כתלות בזמן ההזמנה מראש.*

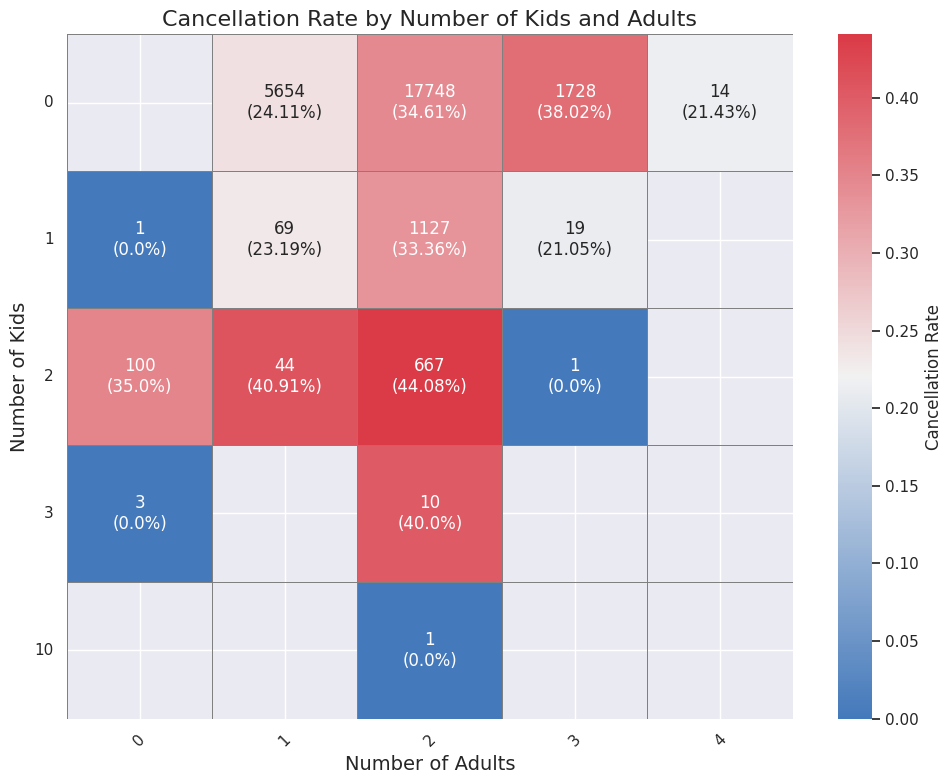


*את ההצגה ההיסטוגרמית לעיל ביצענו גם עבור העמודות בעלות הערכים הקטגורים ביחס לסיכוי לבטל באחוזים*

*נציג לדוגמא את Purchase type . בהן ניתן לראות כי עבור הזמנות מסוג 3,5 ישנו אחוז ביטולים קטן בצורה משמעותית משאר סוגי ההזמנות מה שמאפשר להסיק שידיעת סוג ההזמנה עבור הזמנה עתידית תאפשר לחזות בצורה טובה יותר האם היא תבוטל או לא ולכן ישנה חשיבות שניקח אותו בהמשך בחשבון ובבניית מודלי החיזוי שלנו כנ''ל עבור שאר העמודות שעבורן ביצענו את ההשוואה.*

*בנוסף ערכנו גם טבלת השוואה עבור 3 משתנים, בטבלה זו המצורפת לעיל ניתן לראות השילוב בין כמות המבוגרים לילדים בהזמנה, ואת הסיכוי שלביטול ההזמנה עבור כל אחד מהשילובים, אחוז הביטולים עבור הצירוף מוצג בסוגריים, ומספר המופעים עבור צירוף יחיד מופיע כמספר.*

*לדוגמא: עבור 0 ילדים ו2 מבוגרים אחוז הביטולים עומד על כ 34.61% בעוד כמות הפעמים שמופיע בדאטה האימון הזמנה שמכילה 2 מבוגרים בדאטה הינה כ17,778 .*

*כאן ניתן לראות שעבור שילובים מסויימים של כמו ילדים ומבוגרים שמופיעים בצורה שכיחה דיה בדאטה שלנו יש שוני בין הממוצע ביטולים הכולל של כלל הרשומות לחלק מקומבינציות ( נשים לב כי מספר הקומבינציות הינו משמעותי כאש ישנן מספר לא מבוטל של קומבינציות זניחות עקב מספר נמוך ביותר של מופעים בדאטה שלנו)*

**Data Preprocessing**

* כהכנה ולקראת ההטמעה במודלי הלמידת מכונה עלינו לנרמל את הנתונים בעמודות המספריות ביצענו זאת באמצעות אובייקט **MinMaxScaler** אותו יבאנו מספריית sklearn עבור preprocessing.
* את העמודות הקטגוריות שהעברנו לייצוג מספרי במקום שמי , הגדרנו כINT.

**Data balancing**  למה זה חשוב?

חוסר איזון בכמות הדוגמאות (Class Imbalance):

חוסר איזון בכמות הדוגמאות בין המעמדות השונים יכול להוביל למודל שאינו מתפקד היטב על המעמד המיעוטי. לדוגמה, אם רק אחוז קטן מההזמנות במלון מבוטלות, מודל שלא מתמודד עם חוסר איזון עלול לנבא שהזמנות כמעט ולא מתבטלות, מה שלא נכון בפועל.

שימוש ב-Borderline SMOTE:

טכניקת Borderline SMOTE מייצרת דוגמאות חדשות למעמד המיעוט, בעיקר באזור הגבול (boundary) של המעמד. זה מאפשר למודל ללמוד טוב יותר את התכונות הייחודיות של המעמד המיעוטי ולשפר את ביצועי המודל על נתונים לא מאוזנים.

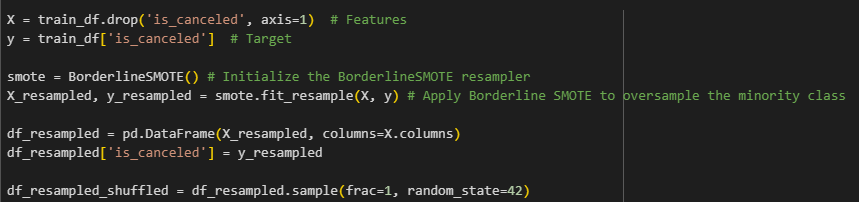
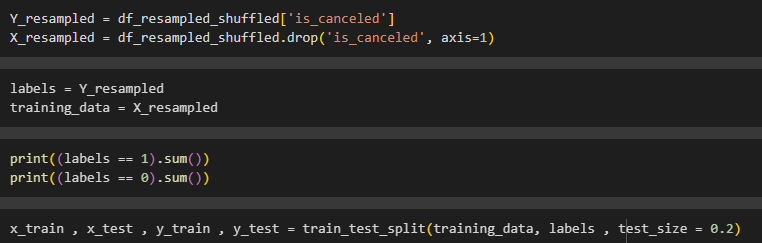
ערבוב הנתונים:

ערבוב הנתונים מבטיח שהנתונים מפוזרים באופן אקראי, מה שמונע הטיות במהלך האימון ומבטיח שהמודל יקבל תמונה נכונה ומלאה של הנתונים.

חלוקה למערכי אימון ובדיקה:

חלוקה נכונה למערכי אימון ובדיקה מאפשרת להעריך את ביצועי המודל על נתונים שלא נראו במהלך האימון, מה שנותן אינדיקציה טובה יותר לביצועי המודל על נתונים חדשים ואמיתיים.

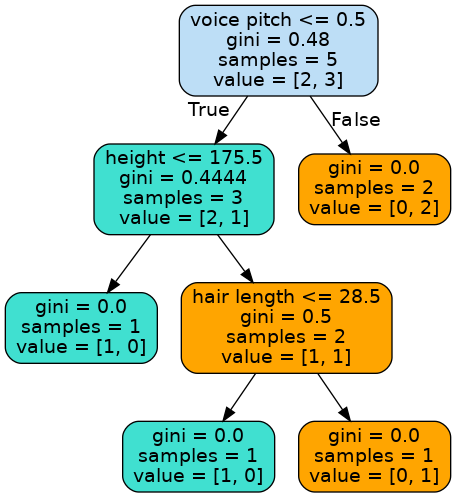
**במקרה שלנו חלוקת היחס בין האימון לבדיקה הינו 80%- 20%**

**כאשר ערך המטרה היינו ה -is canceled .**

**Data Models and Hyperparameters Explanations**

המודלים בהם השתמשנו:

**Decision Tree**

(עץ החלטות) הוא מודל חיזוי שבו הנתונים מחולקים לצמתים על פי קריטריונים מסוימים (כמו gini או entropy).

כל צומת בעץ מייצג פיצול של הנתונים למספר תתי קבוצות, עד שמגיעים לעלי העץ המייצגים את התחזיות הסופיות.

מתחילים משורש העץ ובוחרים את הפיצול הראשון על בסיס קריטריון כלשהו

(למשל, gini או entropy).חוזרים על התהליך עבור כל תת קבוצה שנוצרת עד שמגיעים לעלי העץ או לעומק מקסימלי מוגדר מראש.

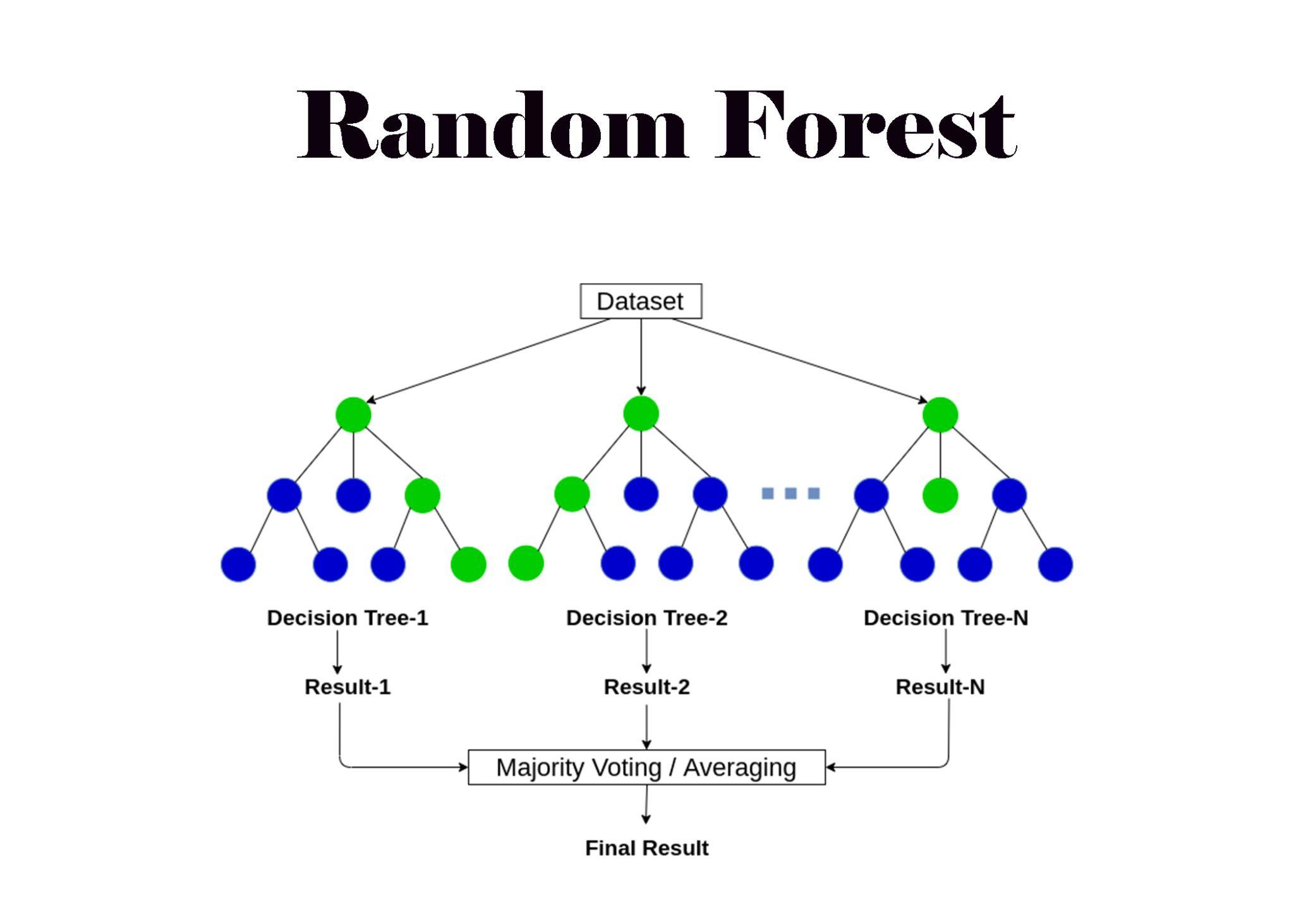
יתרונות:

קל להבנה ולפירוש.

דורש הכנה מינימלית של הנתונים.

מתאים לבעיות סיווג (classification)

**Random Forest** (יער אקראי) הוא מודל המורכב ממספר רב של עצי החלטה.

המודל משתמש בריבוי עצים כדי לשפר את הדיוק ולמנוע overfitting. התחזיות של כל עצי ההחלטה משולבות יחד כדי לתת תחזית סופית.

בונים מספר רב של עצי החלטה על תתי-קבוצות שונות של הנתונים.

משלבים את התחזיות של כל העצים כדי לקבל תחזית סופית (למשל, ממוצע או רוב קולות).

יתרונות:

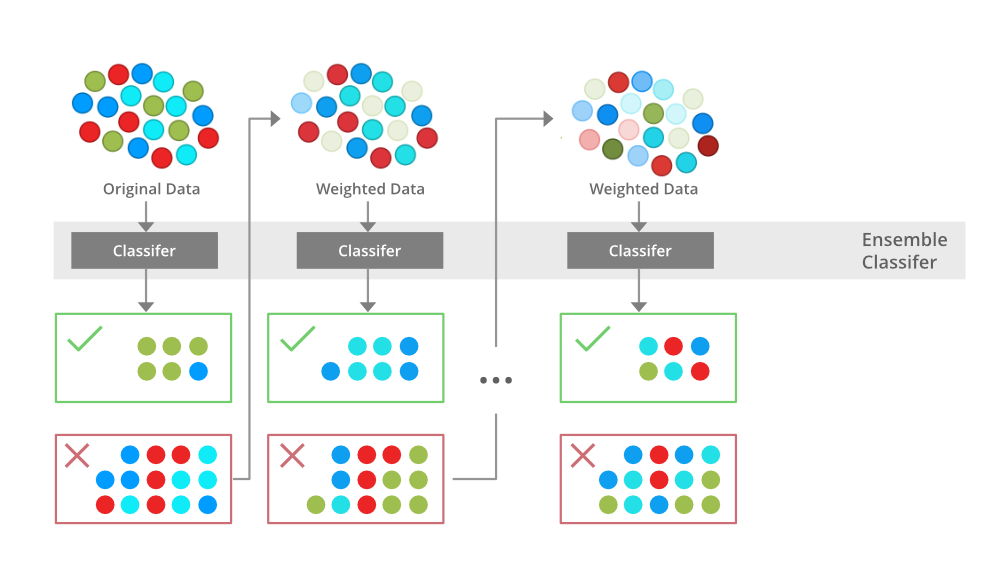
מספק ביצועים גבוהים בהשוואה לעץ החלטות בודד.

פחות נוטה ל-overfitting.

מתאים לטיפול בבעיות סיווג ורגרסיה.

**AdaBoost )Adaptive Boosting)** הוא אלגוריתם Boosting שמשתמש במספר מודלים חלשים (לרוב עצי החלטה קטנים) כדי ליצור מודל חזק. האלגוריתם מתאים משקלים גבוהים יותר לדוגמאות שהמודלים הקודמים נכשלו בחיזוי שלהן.

מתחילים עם מודל חלש ומאמן אותו על כל הנתונים.

מודלים נוספים מאומנים על דוגמאות שהמודלים הקודמים נכשלו בחיזוי שלהן.

כל המודלים משולבים יחד כדי ליצור תחזית סופית.

יתרונות:

משפר את הביצועים של מודלים חלשים.

מתאים לבעיות סיווג ורגרסיה.

**XGBoost )Extreme Gradient Boosting)** הוא אלגוריתם Boosting מתקדם שמשתמש בעצים כמו AdaBoost, אך משפר את הביצועים באמצעות טכניקות שונות כמו קיצוץ (pruning) וניהול זיכרון יעיל.

בונים עצי החלטה באופן איטרטיבי, כאשר כל עץ משפר את התחזיות של העצים הקודמים.

משתמשים במידע מהשגיאות של העצים הקודמים כדי לשפר את התחזיות הבאות.

יתרונות:

מספק ביצועים גבוהים מאוד.

מתאים לטיפול בבעיות סיווג ורגרסיה.

כולל אופטימיזציות רבות שמשפרות את הביצועים ומהירות האימון.

**Model hyperparameters meaning and range explanation**

הסבר על הטווחים וההיפרפרמטרים שנבחרו לכל אחד מהמודלים

**1. Decision Tree**

היפרפרמטרים:

`**criterion**`: הקריטריון המשמש לפיצול הצמתים בעץ. הערכים האפשריים הם 'gini' ו-'entropy'.

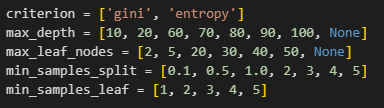
`**max\_depth**`: העומק המקסימלי של העץ. ערכים אפשריים הם מספרים שלמים שונים ו-None (ללא הגבלה).

`**max\_leaf\_nodes**`: מספר העלים המרבי בעץ. ערכים אפשריים הם מספרים שלמים שונים ו-None.

`**min\_samples\_split**`: המספר המינימלי של דוגמאות הדרוש לפיצול צומת פנימי. ערכים אפשריים הם מספרים שלמים שונים ומספרים עשרוניים בין 0 ל-1.

`**min\_samples\_leaf**`: המספר המינימלי של דוגמאות הדרושות להיות בעלה. ערכים אפשריים הם מספרים שלמים שונים.

טווחים שנבחרו:



הסבר לטווח:

`**criterion**`: שתי האפשרויות הנפוצות ביותר לפיצול עצים, מה שמאפשר למודל לבדוק איזה קריטריון מספק ביצועים טובים יותר.

`**max\_depth**`: טווח רחב של ערכים כדי לבדוק את השפעת עומק העץ על הביצועים ולמנוע overfitting.

`**max\_leaf\_nodes**`: בחירת ערכים שונים של מספר עלים כדי לבדוק את השפעתם על הביצועים.

`**min\_samples\_split**` ו-`**min\_samples\_leaf**`: טווחים שונים כדי לבדוק את השפעת מספר הדוגמאות המינימלי לפיצול ועלה על הביצועים. ערכים נמוכים מדי יכולים להוביל ל-overfitting וערכים גבוהים מדי יכולים להוביל ל-underfitting.

**2. Random Forest**

היפרפרמטרים:

`**n\_estimators**`: מספר העצים ביער.

`**criterion**`: הקריטריון המשמש לפיצול הצמתים בעץ.

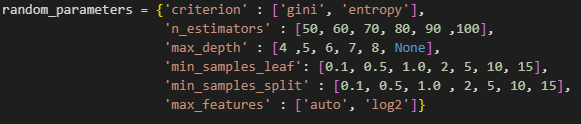
`**max\_depth**`: העומק המקסימלי של כל עץ.

`**min\_samples\_split**`: המספר המינימלי של דוגמאות הדרוש לפיצול צומת פנימי.

`**min\_samples\_leaf**`: המספר המינימלי של דוגמאות הדרושות להיות בעלה.

`**max\_features**`: מספר התכונות המרבי להיחשב לפיצול.

טווחים שנבחרו:



היגיון מאחורי הבחירה:

`**n\_estimators**`: בחירת מספרים שונים של עצים כדי לבדוק את השפעת כמות העצים על הביצועים ולמצוא את האיזון בין דיוק זמן חישוב.

שאר ההיפרפרמטרים דומים לאלו של Decision Tree, אך מותאמים ל-Random Forest כדי לבדוק את השפעתם על יער של עצים רבים.

`**max\_features**`: בחירת מספר התכונות המרבי כדי לבדוק איך שינוי במספר התכונות משפיע על הפיצול ועל ביצועי המודל.

**3. AdaBoost**

היפרפרמטרים:

`**n\_estimators**`: מספר המודלים החלשים (weak learners).

`**learning\_rate**`: קצב הלמידה שמכוון את תרומת כל מודל חלש למודל הסופי.

טווחים שנבחרו:



היגיון מאחורי הבחירה:

`**n\_estimators**`: בחירת מספרים שונים של מודלים חלשים כדי לבדוק את השפעת כמות המודלים על הביצועים.

`**learning\_rate**`: בחירת ערכים שונים של קצב הלמידה כדי לבדוק את השפעתם על הביצועים. קצב למידה נמוך מדי עלול לגרום למודל לא ללמוד מספיק, בעוד שקצב למידה גבוה מדי עלול לגרום ל-overfitting.

**4. XGBoost**

היפרפרמטרים:

`**n\_estimators**`: מספר הסיבובים של boosting.

`**learning\_rate**`: קצב הלמידה שמקטין את תרומת כל עץ כדי למנוע overfitting.

`**max\_depth**`: העומק המקסימלי של כל עץ.

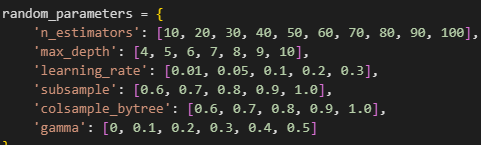
`**min\_child\_weight**`: הסכום המינימלי של משקלי הדוגמאות הדרוש בצומת.

`**subsample**`: אחוז הדוגמאות שמשתמשים בהן לכל עץ.

`**colsample\_bytree**`: אחוז התכונות שמשתמשים בהן לכל עץ.

`**gamma**`: ההפחתה המינימלית בהפסד הנדרשת לפיצול צומת.

טווחים שנבחרו:



היגיון מאחורי הבחירה:

`**n\_estimators**`: בחירת מספרים שונים של סיבובים כדי לבדוק את השפעת כמות הסיבובים על הביצועים ולמצוא את האיזון בין דיוק זמן חישוב.

`**max\_depth**`: בחירת ערכים שונים של עומק העץ כדי לבדוק את השפעת עומק העץ על הביצועים ולמנוע overfitting.

`**learning\_rate**`: בחירת ערכים שונים של קצב הלמידה כדי לבדוק את השפעתם על הביצועים.

`**min\_child\_weight**`: בחירת ערכים שונים כדי לבדוק את השפעת המשקלים המינימליים על הביצועים.

`**subsample**`: בחירת אחוזים שונים של הדוגמאות כדי לבדוק איך שינוי בכמות הדוגמאות שמשתמשים בהן משפיע על הביצועים.

`**colsample\_bytree**`: בחירת אחוזים שונים של התכונות כדי לבדוק איך שינוי בכמות התכונות שמשתמשים בהן משפיע על הפיצול ועל ביצועי המודל.

`**gamma**`: בחירת ערכים שונים של הפחתת ההפסד כדי לבדוק את השפעתם על הפיצולים.

**Models hyperparameters optimization**

ביצוע אופטימיזציה לפרמטרי המודלים אנו בחרנו ב RandomizedSearchCV

למה צריך לבצע אופטימיזציה באמצעות **RandomizedSearchCV** על המודלים הללו?

כאשר אנחנו מאמנים מודל למידת מכונה, ישנם פרמטרים שנקבעים מראש, שנקראים היפרפרמטרים. היפרפרמטרים אלו משפיעים באופן משמעותי על ביצועי המודל. כיוון היפרפרמטרים (Hyperparameter Tuning) הוא תהליך חשוב שנועד למצוא את הערכים הטובים ביותר עבור היפרפרמטרים אלו כדי לשפר את ביצועי המודל.

הסיבות העיקריות לביצוע אופטימיזציה של המודלים באמצעות RandomizedSearchCV הן:

שיפור ביצועים:

בחירת ההיפרפרמטרים הנכונים יכולה לשפר באופן משמעותי את הדיוק, היציבות והיעילות של המודל. מודלים עם היפרפרמטרים מכוונים היטב יכולים להשיג תוצאות טובות יותר בבדיקות ובתחזיות.

מניעת Overfitting ו-Underfitting:

אופטימיזציה של היפרפרמטרים יכולה לעזור למצוא את האיזון הנכון בין מורכבות המודל ופשטותו. מודל מורכב מדי עלול לבצע Overfitting וללמוד את רעשי הנתונים במקום את המגמות האמיתיות. מודל פשוט מדי עלול לבצע Underfitting ולא ללמוד מספיק את הנתונים.

שיפור הכללת המודל:

אופטימיזציה עוזרת למודל ללמוד בצורה טובה יותר ממערך הנתונים וליישם את הידע הזה על מערכי נתונים חדשים שלא נראו במהלך האימון, ובכך לשפר את היכולת של המודל להכליל (generalize) על נתונים חדשים.

חסכון בזמן ובמשאבים:

RandomizedSearchCV מאפשר לבדוק מספר רב של שילובים של היפרפרמטרים בצורה יעילה ומהירה. במקום לבדוק את כל השילובים האפשריים, RandomizedSearchCV בודק מספר קבוע של שילובים אקראיים, מה שמאפשר להגיע לתוצאות טובות בפחות זמן חישוב.

RandomizedSearchCV מאפשר לנו לבדוק מספר רב של שילובים של היפרפרמטרים באופן אקראי מתוך טווח מוגדר מראש. זה נותן לנו מספר יתרונות:

יעילות בזמן חישוב:

RandomizedSearchCV בודק מספר קבוע של שילובים אקראיים זה חשוב במיוחד כשיש לנו הרבה היפרפרמטרים וטווחים רחבים של ערכים.

יכולת לכסות טווח רחב של ערכים:

מכיוון ש-RandomizedSearchCV בודק שילובים אקראיים, יש סיכוי גבוה יותר למצוא שילובים טובים של היפרפרמטרים מבלי לבדוק את כל השילובים האפשריים.

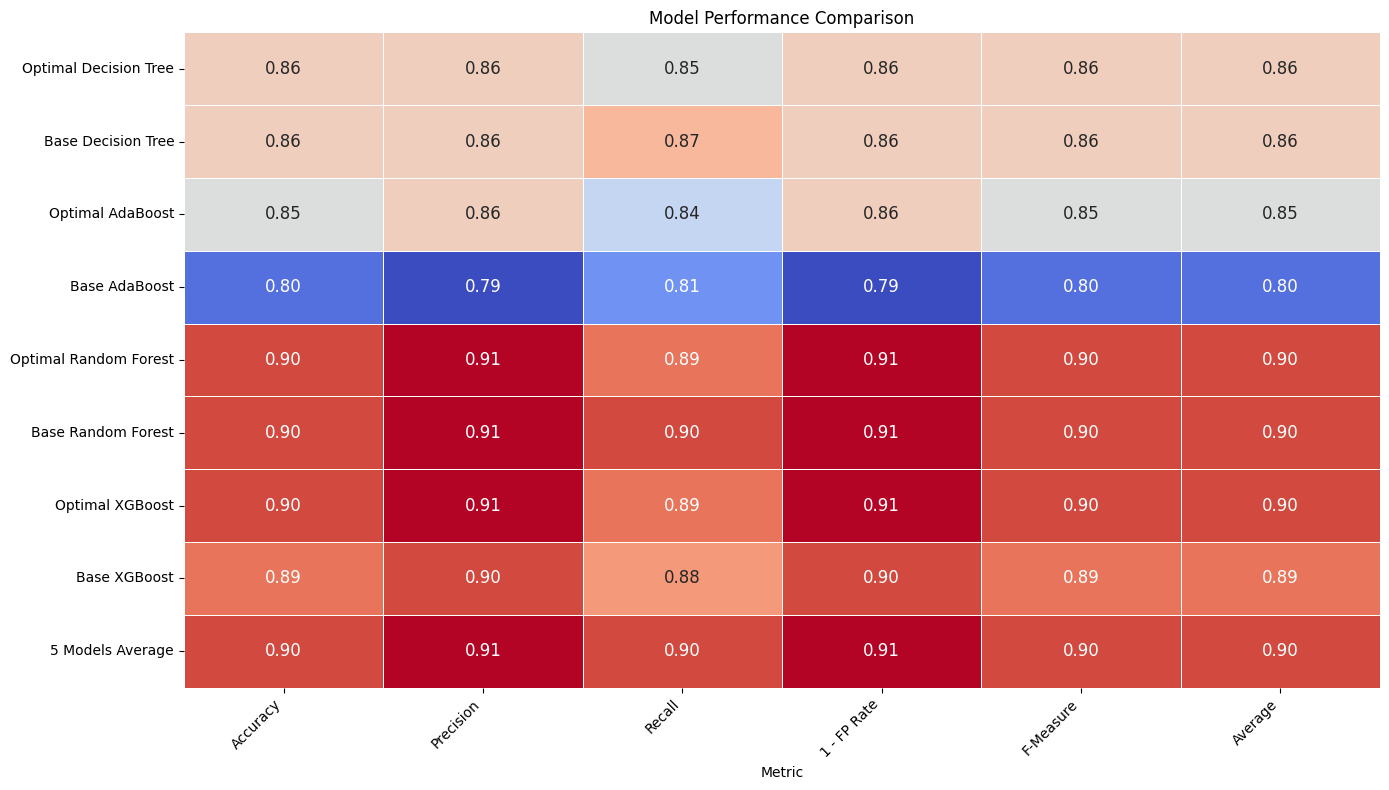
שיפור ביצועים:

RandomizedSearchCV מאפשר למצוא שילובים טובים של היפרפרמטרים שמשפרים את ביצועי המודל מבלי לבצע חיפוש ממצה של כל השילובים האפשריים, מה שיכול להיות בלתי מעשי כשיש הרבה היפרפרמטרים.

**Model Training and preformence Evaluation**

עבור פרויקט חיזוי ביטול הזמנות בבתי מלון, נבחרו מודלים שונים של למידת מכונה כדי להעריך את הביצועים שלהם בחיזוי ביטולים.

עבור כל אחד מהמודלים חישבנו accuracy , presition,recall,fpr,f-score. בהתאם לערכי ה tp,tn,fp,fn שמצאנו:



*סיכום תוצאות המודלים בטבלה כאשר ניתן לראות בצד שמאל בציר הY את המודלים השונים המכילים גרסה לפני אופטימיזציה ואחריה*

*ובציר ה X ניתן לראות את התוצאות ערכי המטרה השונים , כאשר בעמודה הימנית מחושב הממוצע שלהם*

*מכאן אנו רואים כי עלינו לבחור את שני מודלים ה-Random forest , שני מודלי ה- XGBoost ומודל חמישי הBase decision Tree עבור כל רשומה נבצע פרדיקציה עבור כל אחד מהמודלים ונחליט האם ערך המטרה 'Is\_Canceled' שווה אפס או אחד על פי הרוב*

*( הרוב מ5, ז''א 3 ומעלה)*

**Test preprocessing and prediction**

**עבור בחינת הקובץ ה-test**

נשמור את הקובץ כdata frame באמצעות pandas כ - test\_df

נבצע שינויים בדאטה כמו שיבצענו עבור קובץ הtrain

ניתן ערך מספרי למידע קטגורי room\_type,board\_type,purchase\_type

נשלים ערכים חסרים בדומה לדרך שפורטה עבור ה-train

נוריד את עמודת ה-ID

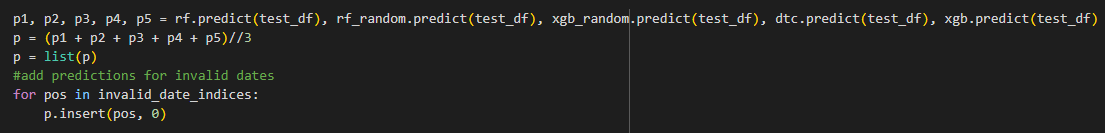
ננרמל את הנתונים המספריים בין 0 ל-1.

עבור העמודות הקטגוריות שהצגנו במספרים נציג אותן במשתנים מטיפוס -int.

שמרנו את הדאטה לאחר הכנה

הרצנו על 5 המודלים וביצענו עבורו חישוב שבו נלך אחרי הרב:

נחלק ב 3 ללא שארית ככה שעבור חיזוי של P=1 עבור 3 מודלים ומעלה נחזה P=1 אחרת נחזה עבור אותה רשומה P=0



נוסיף את העמודה is\_canceled למודל ונשמור אותו כ-test\_with\_predictions .

(\*כלל קטעי הקוד מופיעים במחברת הפייתון וcsv עם התוצאות מצורף בקובץ הזיפ.)