מטלה 1 - תשובות

במטלה התבקשנו לחזות מדדים שונים במסגרת ניתוח מאגר מידע של אתר Agoda.

# מטרה 1

המטרה הראשונה הייתה לחזות האם משתמש יבטל את ההזמנה.  
לטובת כך, פיצלנו את המידע (ביחס של 20% לטסט).  
בעיבוד-המקדים של המידע, המוטיבציה שלנו הייתה לחלץ כמה שיותר מידע מדיד מהנתונים הקיימים: (*מוטיבציה)*

* מילאנו מידע חסר (Nan) בערכי ממוצע
* החלפנו עמודות בוליאניות לאינטג'ריות (0 או 1)
* זמנים:
  + חילצנו מזמן ה-booking את השנה, החודש, היום בשבוע, היום בשנה, שעה *- לזמן בו התבצעה ההזמנה תיתכן השפעה על הביטול (הזמנות בסופי שבוע או באמצע הלילה אולי יתבטלו יותר?)*
  + בתאריך של checkin/out - חילצנו את היום בשנה
  + משך הזמן מה booking עד ל checkin - *יותר ביטולים בהזמנות מוקדמות*
  + משך זמן השהות המיועד - *חופשות ארוכות יותר פחות יתבטלו?*
  + גיל המלון מהכניסה למערכת עד להזמנה - *מלון ישן יבטלו יותר?*
* עלויות ללקוח:
  + עלות כוללת ללילה
  + מחיר חדר ללילה
  + עלות ההזמנה הכוללת פר מבוגר
* מספר ההזמנות שהלקוח ביצע באתר
* ניתוח קוד הביטול: המחיר המקסימלי/מינימלי שהלקוח ישלם במידה ויבטל, האם הלקוח כבר לא יוכל לבטל ללא תשלום - *יותר ביטולים כאשר אין סנקציות*
* משתני דאמי לכל התיוגים הטקסטואלים: סוג האירוח, מאפייני המלון, קוד המדינה של המלון, קודים של האורח וכו'

לאחר ניתוח המידע בנינו מודל ואימנו אותו על הטסט דאטא.  
המודל עוסק לרוב עם משתנים בינאריים, לכן עץ החלטה היה הבחירה האינטואטיבית שלנו.

בבחינה של עומקים שונים היה ניתן לראות שבעומק גדול יותר יש overfit.

# מטרה 2

המטרה השנייה הייתה לחזות את עלות ההזמנה, עבור ההזמנות שעשויות להתבטל.  
העיבוד-המקדים של המידע היה דומה לסעיף הקודם, עם שינויים בהתאם: (*מוטיבציה)*

* ללא מידע על המחירים
* חילצנו מקוד ההזמנה את מספר הימים המקסימלי/מינימלי בקודי הביטול, האם הלקוח כבר לא יוכל לבטל ללא תשלום.

לאחר ניתוח המידע בנינו מודל ואימנו אותו על הטסט דאטא.  
המודל עוסק לרוב עם משתנים בינאריים, לכן עץ החלטה היה הבחירה האינטואטיבית שלנו.

בבחינה של עומקים שונים היה ניתן לראות שבעומק גדול יותר יש overfit.