יום שישי 09 יוני 2023

מטלה 1 - Agoda

"כל המודלים טועים, אך חלקם שימושיים. **בעיקר שלנו**"

מגישים: מירב כהן-גנוז, שלום בלוי, דוד אנגל קלך, שי פריפשטיין

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

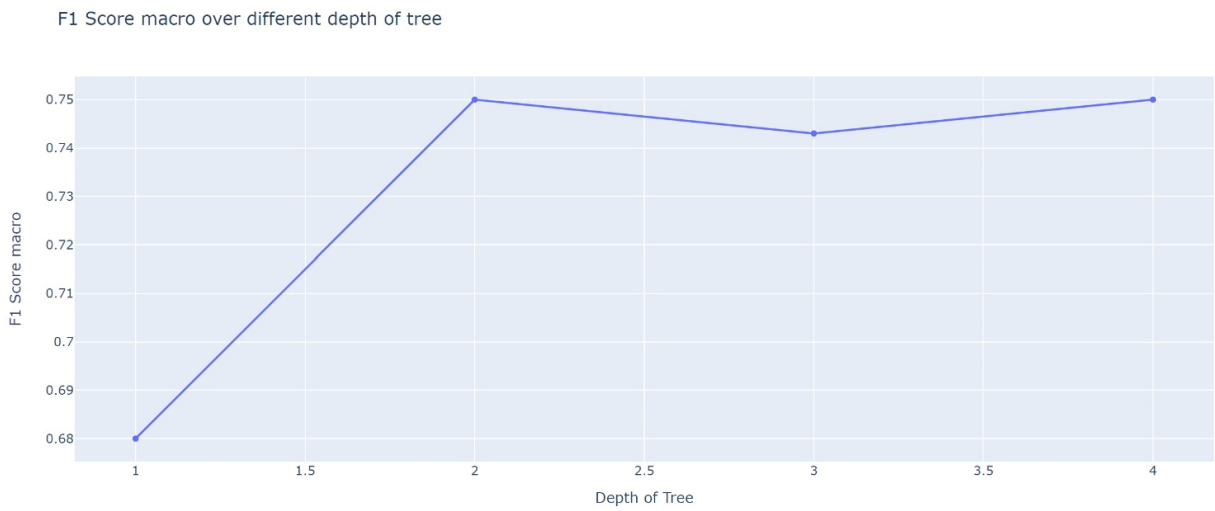
מטרה 1

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 יתבטלו יותר או פחות?*
  - 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 מצאנו שהעומק הטוב ביותר הוא k=2, כאשר בעומק 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.

עבור המודל השתמשנו באלגוריתם הרגרסיה HistGradientBoostingRegressor, הוא אלגוריתם רגרסיה המבוסס היסטוגרמה עם הגברת שיפוע. הוא יעיל, מטפל בערכים חסרים ויכול להתמודד עם מערכי נתונים גדולים. הוא משתמש בטכניקות רגוליזציה ברמה גבוהה. לאחר ניסיון של מספר מודלים החלטנו לבחור במודל זה.

פונקציית main

תחילה יצרנו שני לומדים, אחד קלסיפייר שחוזה את עמודת ה"cancellation_datetime ", והשני רגרסור שחוזה את עמודת ה"original_selling_amount". את המודלים יצרנו מבעוד מועד, ושמרנו אותם בקבצים מתאימים.
עבור המשימה הראשונה השתמשנו במודל הראשון כמו שהוא.
עבור המשימה השנייה, תחילה חזינו את עמודת ה"original_selling_amount" בעזרת המודל רגרסיה, ואז העברנו את הטסט החדש למודל הראשון שיחזה בעזרת קלסיפיקציה את " cancellation_datetime ".
לפי הפורום מילאנו את עמודת ה"original_selling_amount" בערך החסר, כלומר המחיר המקורי שהיה שם.

מסקנות

1) התחושה מידע בתחום היא מתעתעת - צריך לזכור להגיע לידע בצורה נקייה כדי לא להסיק מסקנות שגויות.
2) חלוקת עבודה נכונה היא מפתח לעבודה יעילה, אבל צריך לזכור גם לשתף פעולה לטובת רעיונות טובים ותוצאות מוצלחות.
3) עיבוד מקדים של הדאטא הוא חשוב ביותר ומהווה את ההבדל בין מודל מוצלח למודל פח.
4) היה לנו כיף, מקווים שגם לכם!

הקבוצה