Introduction to Machine Learning (67577) - Hackathon 2023

Hotel Cancellations

**Preprocess** –

* **הסרת** **פיצ'רים** – הסרנו פיצ'רים שלערכיהם אין ערך לחיזוי ביטול הזמנה או מחיר הזמנה, למשל מספר הזמנה. בנוסף, הסרנו פיצ'רים שמייצגים מידע דומה לפיצ'רים אחרים. למשל, הסרנו פיצ'רים שמכילים מידע על המדינה והעיר בה ממוקמים בתי המלון, והשארנו פיצ'ר שמכיל מידע על האזור שבו ממוקם המלון.
* **פיצ'רים** **קטגוריאליים** – הפעלנו את הפונקציה get dummies על הפיצ'רים שערכים קטגוריאליים, ואין אפשרות להמיר אותם לערכים מספריים בעלי משמעות. למשל:
  + Hotel area code, hotel chain code, original payment currency, original payment method, original payment type, accommodation type name, customer nationality, guest nationality name, charge option
* **ערכים חסרים** – חישבנו לכל פיצ'ר בסט האימון ערך ממוצע של הפיצ'ר. לאחר מכן, החלפנו ערכים חסרים בסט האימון ובסט המבחן בממוצעים של אותו הפיצ'ר. עשינו זאת על מנת לא לוותר על דגימות עם ערכים חסרים בסט האימון ובסט המבחן.
* **הוספת פיצ'רים** – במרבית הפיצ'רים שמכילים תאריך, המידע המשמעותי בעיננו הינו הפרש הזמנים בין עמודת תאריכים אחת לאחרת. למשל, תאריך הצ'ק אין ותאריך הצ'ק אאוט פחות משמעותיים בעינינו לפרדיקציה. לעומת זאת, משך השהייה במלון, כלומר הפרש הימים בין הצ'ק אין לצ'ק אאוט, מהווה מידע מעניין לפרדיקציה. לכן, הסרנו לחלוטין את עמודות התאריכים, ויצרנו מהם עמודות פיצ'רים חדשות, שמכילות את כמות הימים בין התאריכים השונים. פיצ'רים חדשים שהוספנו לדוגמא:
  + משך השהייה – ההפרש בין תאריך הצ'ק אאוט לתאריך הצ'ק אין.
  + משך הזמן שהמלון נמצא באתר – ההפרש בין תאריך הצ'ק אין לתאריך העלאת המלון לאתר.
  + משך הזמן בין ההזמנה להגעה למלון – ההפרש בין תאריך הצ'ק אין לתאריך ביצוע ההזמנה.