יסודות מדעי המידע – פרויקט קורס 2024

את הפרויקט יש לבצע בקבוצות של 3 סטודנטים. במידה ולא תמצאו קבוצה פנו אלי ואני אשדך ביניכם. יום לפני תאריך ההגנה יש להעלות למודל (בתיבה שנפתחה לכך) את מחברת הפרויקט.

בפרויקט זה נחקור התנהגויות מחירים של אתרי Booking.com ו Expedia.com. שימו לב שאת עיקר הדרישות ניתן כמובן לממש באמצעות החומר אותו למדתם. יחד עם זאת, בדרישות הפרויקט שולבו גם מספר משימות המתייחסות לחומר שעליכם ללמוד/להתנסות לבד. מעבר לדרישות עצמן – הרגישו חופשיים לקיים ניסויים עם שיטות נוספות על מנת לשפר את התוצאות – זה רק יכול להוסיף.

הסבר קצר לגבי טרמינולוגיה של עולם המלונות:

TTT = Time to Travel – כלומר הפרש הימים בין תאריך החיפוש באתר לתאריך ה checkin לחדר המבוקש.

LOS = Length of Stay – כלומר מספר הלילות שנרצה ללון במלון.

Snapshot Date = התאריך בו ביצעתם את דגימת האתרים.

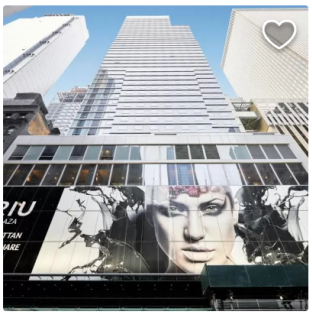
הגשת הפרויקט תעשה באמצעות מחברת Ipynb. שימו לב שעליכם לפרט (עם הערות) את כל הניסויים שביצעתם במחברת – גם את מה שהיה "בדרך" אל הפתרון. כך גם במקרה שלא הצלחתם לבצע סעיף מסוים, נדע מה ניסיתם.

בהמשך יתבצעו הגנות בקורס (במועדים שיתפרסמו). בהגנה עליכם להציג את המחברת ולענות לשאלות שתישאלו. השאלות מתייחסות בעיקר למה שביצעתם בפרויקט (בין אם התבקשתם ובין אם בחרתם/יזמתם) אך לא רק – אלא גם לחומר שנלמד במהלך הסמסטר, שעליכם לשלוט בו לקראת ההגנה. כל סטודנט החבר בקבוצת הפרויקט צריך להכיר היטב את הקוד שנכתב על ידי כל אחד מחברי הקבוצה.

שלב א' – Scraping:

השתמשו בספריה המבצעת Web scraping מתוך פייתון (לבחירתכם. לדוגמא: Selenium, Beautifulsoup, Scrapy) על מנת לדגום את תוצאות החיפוש על פי הקרטריונים הבאים:

- ניו יורק
- 2 מבוגרים, 0 ילדים
- חדר אחד
- תאריכים: כל הקומבינציות של **TTT** בין 1 ל 30 ושל **LOS** בין 1 ל 5 (כלומר סה"כ 150 חיפושים של תאריכי לינה שונים עבור Snapshot מסוים) עבור 3 תאריכי **Snapshot** שונים (סה"כ לפחות 450 חיפושים). יש לקחת בחשבון שייתכן ותצטרכו לתקן באגים בקוד ה scraping ולכן הערכו בהתאם וסיימו את שלב זה כמה שיותר מוקדם על מנת לאפשר מספיק זמן לסריקות עצמן.
- עבור כל אחת מהקומבינציות של TTT ו LOS עליכם לשמור בקובץ (אפשר CSV) את הפרטים המופיעים על הדף עבור 100 בתי מלון לפחות. (שימו לב שבאתר Booking עליכם לעבור בין דפים שונים, ובאתר Expedia עליכם לגלול מטה עד שתקבלו את מספר בתי המלון הנדרש). יודגש שהקוד של ה scraper צריך להיות אוטומאטי לחלוטין ללא התערבות שלכם במהלך הריצה.
- עבור כל מלון עליכם לאסוף את כל השדות המופיעים בתא של המלון המופיע על המסך (כולל שם המלון, ציונים למיניהם, מספר כוכבים/דרוגים, מרחק ממרכז העיר, סוג חדר, מדיניות ביטולים, מחיר וכד'):



Riu Plaza Manhattan Times Square ★★★★

Manhattan, New York · Show on map · 1 km from center · Subway Access

Excellent 8.6
5,811 reviews
Location 9.5

35% off | Early 2023 Deal

Deluxe Family King
2 beds (1 king, 1 sofa bed)
FREE cancellation • No prepayment needed
You can cancel later, so lock in this great price today!

3 nights, 2 adults
~~₪ 2,874~~ ₪ 1,868
+₪ 652 taxes and charges

See availability

שימו לב ששדות מסויים לא בהכרח יופיעו בכל האתרים – למשל מרחק ממרכז העיר יופיע ב Booking אבל לא ב Expedia

- שימו לב שהנתונים שעליכם לאסוף מופיעים רק בדפי החיפוש – אין צורך להיכנס לפרטים של כל מלון על מנת לאסוף נתונים נוספים. במידה וכן ובחרתם כן לאסוף נתונים מדפי המלון עצמם – ציינו זאת בהגנה ובצעו את הניתוחים המופיעים בהמשך בכל פעם בלי ועם הנתונים שהוספתם.

שלב ב' – Exploration + Data preprocessing:

- יש ליצור גרפים של התפלגויות של:
  - מספר ה reviews
  - ציון המלון ב reviews
  - המחירים
- עליכם להסיר מהנתונים בתי מלון שעל פי המחיר מהווים outlier על פי שיטת Tukey (1.5IQR)
- המירו משתנים בעלי אופי Oridinal למספרים בעלי יחס סדר
  - (למשל Good<VeryGood<Excellent)
  - דוגמא נוספת - את סוגי המיטות לפי יחס סדר (ראשית עליכם לחשב את רשימת כל סוגי המיטות (distinct) מתוך כלל הנתונים שבידיכם). לדוגמא King Room אמור לקבל ערך גבוה יותר מ Queen Room
- המירו משתנים בעל אופי Nominal או קטגורי באמצעות מספר עמודות או בשימוש ב One-hot encoding
  - למשל מדיניות cancellation ו prepayment, שם השכונה בניו יורק
  - ניתן להעזר ב https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
- הציגו PairGrid (של ספריית seaborn) עבור המשתנים וכתבו בפיסקה מסקנות/תובנות מהגרף. כמו כן התייחסו לקשר בין המשתנים השונים X לבין משתנה המטרה (מחיר)

שלב ג' – תחזית מחירי החדרים

בסעיף זה עליכם לממש במחברת פתרון לחזות את מחירי חדרי המלון על פי הנתונים שאספתם לגבי כל מלון בשלבים הקודמים.

- שימו לב שעליכם להתנסות במספר אלגוריתמי רגרסיה (לפחות את אילו שנלמדו: LinearRegression, DecisionTreeRegressor, GaussianProcessRegressor) ובנוסף עליכם לבחור עוד 3 אלגוריתמי רגרסיה שלא נלמדו מתוך ספריית sklearn.
- עליכם לוודא שחילקתם את הדאטה ל Train, Test בחלוקה של 30-70 לטובת ה Train בבחירה אקראית.
- עבור כל אלגוריתם עליכם לבצע התנסויות בפרמטרים שונים (למשל עבור עץ רגרסיה לנסות הגדרות שונות לעומק העץ, לבחינת קריטריון החלוקה וכד'. עבור תהליכים גאוסיאנים להתנסות עם קרנלים שונים.

- הציעו פיצ'רים נוספים (המחושבים על סמך הקיימים) ובידקו כיצד הם משפיעים על תוצאות התחזית שלכם. למשל אולי כדאי לקחת לא רק את ערכי ה TTT אלא גם את היום בשבוע של ה checkin, את הקירבה לסוף החודש, קירבה לחג בתקופה וכד'
- עבור לפחות 2 אלגוריתמים - דרגו את המשתנים השונים על פי מידת השפעתם על תוצאות הרגרסיה. הציעו לפחות 2 דרכים שונות לבצע זאת (נקרא גם feature importance).
  - דרך אחת תתבסס על הדרך בה האלגוריתם הספציפי עובד
  - דרך אחרת תתעלם מאיך האלגוריתם עובד ותתייחס אליו כקופסה שחורה
- עבור כל הרצה של אלגוריתם, חשבו את השגיאה על ה Test ועל ה Train והדפיסו Residual Plot כפי שביצענו בכיתה. כתבו פסקה מסכמת לגבי התובנות מהגרף.
- מדדי שגיאות נדרשים עבור כל אלגוריתם: RMSE, MSE, MAE ,R2.
- נתחו את התוצאות שקיבלתם באלגוריתמים השונים. התייחסו לחולשות/יתרונות של אלגוריתמים שונים שהשתמשתם בהם ונסו להסביר מה הסיבה לכך (במיוחד המתקדו במקרים בהם אלגוריתם מסוים קיבל תוצאות טובות/גרועות משמעותית יחסית לאלגוריתמים אחרים).
- התנסו בשיטות נרמול שונות על הנתונים (לא חייבים על כל השדות) טרם הרצת האלגוריתמים ודונו בהשפעם השיטות על ביצועי האלגוריתמים השונים
- הפרידו את הניתוחים בין הנתונים שהורדתם מ Booking לבין אילו של Expedia. מצאו הבדלים בין המודלים על האתרים השונים, לא רק בביצועים אלא גם במשתנים המשפיעים על הפרדיקציות, חוזקות/חולשות אלגוריתמים מסוימים וההבדלים הניכרים לעין
- הציגו גרף המראה את התפלגות השגיאות R2 של המודל הטוב ביותר שלכם על פני המלונות השונים
- בצעו את הניסוי שוב על האלגוריתם הטוב ביותר שקיבלתם – הפעם על חלוקה שונה של Train ו Test :
  - ה Train יכיל את הנתונים עבור TTT<=25
  - ה Test יכיל את הנתונים עבור TTT>25 (כך שלעשה אנו בודקים כאן את היכולת לחזות את המחירים ב"עתיד"

שלב ד' – למידת פערי מחירים באתרים מתחרים

- בנו מודל החוזה את הפרש המחירים בין Booking ל Expedia (שימו לב שעליכם לבצע ראשית הצלבה של בתי המלון המשותפים בלבד) בהינתן הנתונים המופיעים על המסך מ Booking ו Expedia גם יחד. בדומה לדרישות לעיל בסעיף הקודם – גם כאן התייחסו לדרישות הטכניות (חלוקת Train/Test, הרצת מספר אלגוריתמים, בחינת השגיאות וכד')

שלב ה' – ביצוע Reverse Engineering לאלגוריתמי המיון של האתרים

- שימו לב כי בברירית המחדל (מבלי שתבחרו באופציית המיון באתר) - האתרים מציגים תוצאות בסדר מסוים על פי שיקולים שונים. המיקום של המלון בדף התוצאות עשוי להיות מושפע למשל מהדירוגים של המלונות, מהמיקום, ממספר ה reviews, ועוד. עליכם לתכנן ולממש פתרון שיאפשר (רק על סמך הנתונים שהורדתם בסעיף הראשון) למיין בעצמכם דף תוצאות. על מנת לבצע זאת עליכם להשתמש בשיטות למידה שלמדתם בקורס.
- על מנת לבדוק את ביצועי אלגוריתם המיון שבניתם, נתוני ה train צריכים להכיל את כל הנתונים מה TTT הזוגיים, ונתונים ה Test יכילו את כל הנתונים מה TTT האי-זוגיים.
- עבור בדיקת test של אלגוריתם המיון שבניתם, עליכם ראשית לבצע shuffle אקראי על נתוני ה Test ואז לתת לאלגוריתם שבניתם למיין מחדש. את גודל השגיאה של האלגוריתם שלכם על דף נתון תממדו באמצעות סכום ריבועי הפרשי האינדקסים של תוצאות החיפוש על דף נתון. לאחר מכן יש להציג את ממוצע וסטיית התקן של השגיאות שנתקבלו מכל דפי ה Test

ב ה צ ל ח ה !