מטלה שלישית - חיזוי עתיד
כלכלה בעולם ה Big Data
ד"ר רועי ששון; אופיר בצר

להגיש לכל המאוחר עד:
2022-08-31 23:59
לא יתקבלו הגשות באיחור.

פרטים חשובים על המטלה:

1. יש לבצע את המטלה **בזוגות**. לא יבדקו מטלות באופן יחיד, במידה ומתעוררת בעיה בנושא אנא הודיעו וימצא פתרון.
   a. רק אחד מחברי כל קבוצה צריך להגיש את המטלה במודל.
   b. נא לציין באופן ברור מי הם חברי הקבוצה בתוך הדוח ובשם של הקובץ של הדוח באופן הבא:
      EinBDW22_A3_{name #1}_{name #2}.pdf
      לדוגמא:
      EinBDW22_A3_roy_sasson_ophir_betser.pdf
2. משקל המטלה הוא **70%** מציון הקורס.
3. את כל הניתוחים יש לבצע ב **R בלבד**.

יש להגיש:

1. דו"ח מפורט ובו התשובות שלכם על השאלות, בצירוף גרפים וטבלאות רלוונטים. על הדו"ח להיות ממוקד - אין להגיש יותר מ 6 עמודים, וינתן משקל על אסתטיות של הגרפים ו/או הטבלאות המצורפים בו.
   a. יש לצרף לכל אחד מהגרפים כותרות המסבירות את תוכנו.
   b. יש לוודא שהגרפים שאתם מצרפים קריאים, וברורים, מובנים ואסתטיים.
   c. אין צורך לכתוב הרבה מלל, אלא לענות באופן קצר וממוקד.
   d. פורמט הקובץ צריך להיות **pdf**.
   e. שפת הדוח צריכה להיות **עברית**, גופן דוד 12, רווח בין השורות 1.5.
2. קובץ csv בפורמט הבא: קובץ () אשר בו יוגשו תחזיות המודל שלכם.
   a. על הקובץ להיות בדיוק בפורמט המצורף - אותם שמות עמודות ואותם ערכים בעמודת ה partition_datetime.
   b. על שם קובץ להיות בפורמט הבא: EinBDW22_A3_predict_{name #1}_{name #2}.csv
3. **יש** לצרף את סקריפט ה R הכולל את הניתוחים שביצעתם.

על סט הנתונים:
במטלה זו תבצעו ניתוח וחיזוי של מחירים הנתיב המהיר.
מחירי הנסיעה בנתיב מעודכנים בהתאם לביקוש הנהגים, כך בשעות העומס מחיר הכניסה לנתיב המהיר עולה.
מחיר הכניסה לנתיב מעודכן באתר האינטרנט, וכך ניתן לעקוב על השינויים בעלויות הנסיעה בו.

את המחיר העדכני אנו אוספים בעזרת זחלן רשת. טבלאות עדכניות יפורסמו מיידי שבוע בדרייב (price_data_yyyymmdd). בכל טבלה יהיו 2 עמודות בלבד: זמן הדגימה (POSIXt) ומחיר הנסיעה הממוצע (numeric).

שימו לב שיתכנו רשומות חסרות בסט הנתונים, קחו זאת בחשבון בעת הניתוח וניקוי הנתונים. למתעניים, זהו סקריפט הזחלן שאוסף את הנתונים (python).

שאלות מחקר: (חלקים 1 עד 4)

חלק 1 - אנליסט המוצר (45%)
חלק זה של המטלה כולל את ניתוח הנתונים המרכזי במטלה.
עליכם "לספר את הסיפור שמאחורי הנתונים" באופן שלוקח בחשבון את משימת התחזית בחלק הרביעי של המטלה.

הכנת וניתוח הנתונים כולל את השלבים הבאים:

1. טעינת הנתונים
2. סידור וניקוי שלהם
3. הוספת משתנים מסבירים (לדוגמא משתנים מבוססי תאריך, מזג אוויר, נתונים קורונה או כל מידע אחר שלדעתכם רלוונטי).
4. ביצוע EDA

חלק 2 - הכלכלן (10%)

בחרו משתנה מסביר אחד מהרשימה, או אחר שאתם חושבים שיכול להיות מעניין: [**מזג אוויר, סוג היום**]. דונו ותארו כיצד ניתן לגלות מה הקשר הסיבתי בינו לבין מחירי הנתיב המהיר מבחינה כלכלית. בניתוח תארו כיצד הייתם ממליצים לבודד את ההשפעה הסיבתיות. כחלק מהניתוח עשו שימוש במודל קלאסטרינג k-means ופרשו את תוצאותיו.

חלק 3 - האנליסט העסקי (5%)

האם הנסיעה בנתיב המהיר משתלמת?

ישראל ישראלי מתגורר בירושלים ועובד בתל אביב. הוא מרוויח שכר שעתי של 100 ש״ח, ויש לו את האפשרות לצאת מביתו בשעות 7,8,9,10 או 11 (שעות עגולות בלבד) ברכבו הפרטי. נתחו את הכדאיות של הנסיעה בנתיב המהיר עבור מר ישראלי - **כמה זמן של נסיעה** על הנתיב לחסוך לעומת אלטרנטיבת הנסיעה בנתיב רגיל על מנת שהשימוש בו יהיה משתלם?

חלק 4 - מדען הנתונים (40%)

חלק זה הינו תחרותי בין הקבוצות. עליכם לתת תחזית בעלת mse נמוך ככל הניתן עבור נקודות הדגימה אשר נמצאות הקובץ הזה (בין ה 01-09-2022 עד ה 13-09-2022, בשעות 06:00 עד 12:00 מדי יום).

כחלק מהניתוח, תארו את אופן בניית המודל, אופן האימון שלו ודרך הבדיקה שלכם את תוצאותיו (למשל - האם יש overfitting).