מבוא למדע הנתונים – תרגיל 2

מטרת התרגיל היא ליישם ולשלב ביעילות טכניקות אנליטיות שכוסו במהלך הקורס כדי לטפל בבעיית החלטה עסקית. **ציון התרגיל ידורג באופן יחסי בין חברי הכיתה על סמך ביצועי מודל החיזוי בלבד.** ממוצע הציונים יהיה 80.

שימו לב! אין להשתמש במודלים סטטיסטים שלא נלמדו במסגרת הקורס. מותר ורצוי להשתמש בחומרים שנלמדו באמצעות השקפים, המצגות המוקלטות והחומר הנלווה שצורף לקורס. מותר לשלב מספר מודלים כדי להגיע לפתרון הבעיה.

<u>הנכם נדרשים להגיש:</u>

1. קוד R מתועד היטב
2. קובץ csv עם תוצר החיזוי כמופרט בהמשך

<u>תיאור הבעיה העסקית</u>

השנה בה מתרחש הסיפור היא 2025, החודש – אוקטובר.

ארגון Cross Israel הינו ארגון חדש ללא מטרות רווח האוסף תרומות מאנשים פרטים ע״מ לסייע לאנשים שנקלעו למשבר כלכלי בשל בעייה בריאותית. מידי חודש, נציגי הארגון פונים לאזרחים הרשומים במאגר הארגון ומבקשים מהם לתרום סכום כסף כראות עינהם. הפניה נעשית באופן טלפוני.

בשנה האחרונה, הארגון נהג לפנות ל-1000 אנשים בכל חודש לבקשת תרומה. עם זאת, סך התרומות מאנשים אלו היה נמוך יחסית, וכסף רב בוזבז על מאמצי טלמרקטינג של נציגי הארגון שכוונו לאנשים שבחרו לא לתרום. לפיכך, החליט הארגון לשנות גישה, ולייעל את מערך הטלמרקטינג באמצעות מודלי data science. בנוסף, הוחלט להגביל את מספר שיחות הטלמרקטינג, לשם צמצום עלויות, ולפנות ל-90 איש בכל חודש בלבד.

ברכות! נבחרתם לייעץ לארגון בבחירת 90 תורמים פוטנציאלים מתוך 1000 אליהם כדאי לפנות בחודש אוקטובר. שימו לב: תורם פוטנציאלי שלא מקבל פנייה טלפונית מהארגון לא תורם כלל. לרשותכם שני קבצי מידע:

1. קובץ תרומות בחודשים ינואר עד ספטמבר 2025 - cross_2022C.csv
2. קובץ המכיל פרטי תורמים פוטנציאלים לחודש אוקטובר 2025 – holdout_2022C.csv, מתוכם יש לבחור 90איש, אליהם יתקשרו נציגי הארגון.

בנו מודל (או מודלים) שיאפשרו לבחור את 90 התורמים הפוטנציאליים מתוך 1000 האנשים בקובץ holdout_2022C.csv, עבורם סך התרומה החזוי יהיה מקסימלי, ואליהם תמליצו לארגון Cross Israel לפנות.

<u>שיטת שמירת הקבצים ובדיקתם:</u>

לתרגיל מצורפים שני קבצי R (סקריפטים):

1. Script.R- זהו קובץ העבודה שלכם, בקובץ זה עליכם לטעון את הנתונים להריץ את המודלים ולהפיק את התחזיות. שימו לב שבחרנו עבורכם שמות של חלק מהאובייקטים, אנא היצמדו לשמות הללו.
2. Compiler.R – זהו קובץ שבודק שהקוד שכתבתם רץ ומפיק קובץ CSV כנדרש.

אתם מתבקשים לשמור את קבצי הנתונים וקבצי ה-R בתיקייה אחת, ולא צריכים להגדיר Working Directory. בתחילת קובץ הscript פונקציה שמגדירה את ה-WD להיות התיקייה שבה הקובץ נמצא.

<u>מה להגיש?</u>

יש להגיש את קובץ הסקריפט וקובץ csv בפורמט הבא:

1. מספיקה הגשה אחת לכל קבוצה, אין צורך להגיש בנפרד.
2. אין לשנות את שמות הקבצים:
   - הקוד צריך להופיע בקובץ script.R
   - ההמלצות צריכות להופיע בקובץ csv שמכיל את ת"ז של חברי הצוות

   קוד ה-R מכיל פקודה ששומרת את ההמלצה שלכם בקובץ csv ששמו מכיל את ת.ז של המגישים. (כלומר, אם התז של חברי הצוות הם: 123,234,345 אז הקובץ יישמר 123_234_345.csv)
3. בקובץ ה-csv תהיה עמודה בודדת, שכותרתה: recommendation (באותיות קטנות!!!)
4. תוכן העמודה יהיה ה- donorId של 90 האנשים שאתם ממליצים לפנות אליהם

בנוסף, יש להגיש קובץ R עם הפתרון. שימו לב:

1. ודאו כי ניתן להריץ את הקובץ מתחילתו ועד סופו ללא קבלת errors
2. ודאו כי קובץ ה-R מייצר קובץ CSV כנדרש, וודאו שקובץ ה-CSV מכיל עמודה 1 ובה ה-donorid. אין לצרף את התחזית שלכם את המשתנים הבלתי תלויים האחרים של התורם וכו'.

<u>שיטת ניקוד</u>

הניקוד ייתן על סמך סך התרומה שיתקבל מהתורמים עליהם המלצתם בלבד, ולא על בסיס איכות הקוד!
הציון יהיה יחסי לשאר חברי הקבוצה, והממוצע יהיה 80. כלומר: זוהי תחרות פרדיקציה! אין לשתף פעולה עם סטודנטים שאינם בצוות שלכם.

שימו לב:

- אנחנו מריצים את הקוד שלכם, באמצעות קובץ ה-compiler, אנא וודאו שקובץ זה אינו מייצר שגיאות ובידקו שטבלת ה-csv מכילה את הערכים שלה אתם מצפים. קוד R המייצר error יקבל 0!
- קוד R שלא ייצר כפלט את קובץ התורמים המומלצים שהוגש יקבל 0!
- קובץ ה-csv שייבדק על-ידנו הוא זה שיופק באמצעות ה-compiler, ולא זה שתגישו ידנית.
- הגשות בפורמט שאינו csv יגרמו להורדה של 10 נקודות מהציון.
- שיום (naming) לא נכון של עמודת recommendation יגרום להורדה של 5 נקודות מהציון.
- קבצים שיכילו יותר מ- 90 שורות המלצה יגרמו להורדה של 5 נקודות מהציון לכל שורה מיותרת. בנוסף, נשקלל רק את 90 הרשומות הראשונות.
- קבצים עם פחות מ- 90 איש לא יגרמו להורדה בציון, אך כמובן אינם אופטימלים בהיבט הפתרון.

בהצלחה!!!