מטלה 1 : מבוא למדעי הנתונים

כללי:

- מטרת הפרויקט היא לאפשר תרגול מעשי של עקרונות שיטת DM-CRISP לניהול מחזור החיים הכולל של פרויקט אנליטיקה עסקית בסביבת python.
הפרויקט כולל תרגול של מספר מודלים בכריית נתונים כפי שנלמדו בכיתה ונועד לאפשר חשיפה לכל שלב בשיטה בצורה מעמיקה ויסודית.
- הפרויקט מחייב תיעוד מפורט של הפעולות אשר נעשו בכל שלב לטובת בדיקה איכותית וקבלת ביקורת בונה ומלמדת.
- במשימה זאת תשתמשו בבסיס נתונים הכולל מידע על שיווק מוצר בנקאי (חיסכון) ללקוחות והאם הלקוחות רכשו את המוצר או לא.
בסיס הנתונים והסברים על הפיצ'רים השונים נמצאים במודל בתיקיה של משימה 1.

קווים מנחים:

- יש להגיש מחברת jupyter מסודרת הכוללת כותרות, הסברים מילוליים ותיעוד של הקוד.
- שאלות על הפרויקט ייענו בפורום בלבד
- בבדיקה יינתן דגש על יעילות נראות הקוד והחשיבה האנליטית שלכם במהלך התרגיל, שימו לב!
- זמן הריצה של המחברת כולה לא יעלה על שעה.
- השתמשו בתרשימים וגרפים בשלבים השונים של התרגיל.

הגשה:

- מועד הגשה 30/12/2019
- הפרויקט יוגש בזוגות , יש להירשם לקבוצות במודל.
- מייל עוזר ההוראה : mr.marudi@gmail.com

בהצלחה!!!

הפרויקט:

1. טעינת בסיס הנתונים:

i) הורידו את קובץ הנתונים bank.csv מהמודל לתיקייה מקומית על המחשב שלכם.
ii) קובץ הנתונים ומחברת הjupyter צריכים להיות באותה התיקייה על מנת לאפשר קריאה נוחה של הקובץ בעת ההרצה והבדיקה.
iii) קראו את קובץ הנתונים במחברת העבודה לתוך אובייקט בשם data.

2. Data exploration:

i) השתמשו בפקודה שמציגה מספר מוגבל של השורות הראשונות. בדקו שהפונקציה שהשתמשתם מראה את הנתונים בצורה טובה ובכך תבינו האם טעינת הקובץ עברה בהצלחה.
ii) הציגו את שמות כל העמודות ואת סוג הנתונים שכל אחת מהן מכילה בעזרת פונקציה אחת
iii) הציגו סיכום של כל העמודות הנומריות (ממוצע, ס"ת , חציון מינימום ומקסימום וכו...) מומלץ לבצע זאת בעזרת פונקציה אחת.
iv) עבור העמודות הקטגוריאליות:
   1. עמודת 'y' : המירו עמודה זאת לעמודה בינארית (1, 0)
   2. עמודת 'month' - המירו ל4 עמודות רבעונים בינאריות 1 האם החודש נופל ברבעון 0 אחרת : Q1,Q2,Q3,Q4
   3. המירו את יתר העמודות על ידי שימוש בשיטות שנלמדו בשיעור או מהספרות – הסבירו במספר מילים מדוע בחרתם בכל שיטה.
   <u>המלצה</u> - נסו למצוא קשר בין עמודה לתגית וצרו פי'צרים חדשים, לדוגמא במידה ותמצאו שיש לימים שונים השפעה דומה חברו אותם לפיצר אחד.

v) בדקו אם ישנה קורלציה בין המשתנים השונים, במידה וכן הורידו את העמודות המיותרות הציגו את מטריצת הקורלציה בצורה ברורה.

שאלה: מדוע נרצה להוריד עמודות אלו?

3. Missing values:

i) סיפרו את כמות השורות בהן יש ערכים חסרים והציגו אותם.
ii) עבור כל עמודה עם ערכים חסרים החליטו אם להוריד שורות אלו או השלימו את הערכים לפי מה שנלמד בכיתה (ניתן ואף רצוי להשתמש בשיטות שונות לפיצ'רים שונים). **עבור כל עמודה הסבירו את החלטתכם**.

4. Data normalization:

i) צרו תרשים box_plot והסבירו מה ניתן ללמוד ממנו
ii) האם יש צורך לנרמל את הדאטה? במידה וכן בצעו זאת בשיטה לבחירתכם.

5. Outlier detection:

i) השתמשו בDBSCAN על מנת לחלק את הדאטה לקלסתרים והוציאו כחריגים את כל הדגימות הרועשות (ציון 1-). שימו לב יש למצוא את הפרמטרים האופטימליים למודל כך שמספר הדגימות הרועשות לא יהיה גדול או קטן מדי. ניתן להוסיף את תוצאת המודל כפיצ'ר עבור מודל החיזוי .
ii) מה המשמעות של ריבוי קלסתרים במודל זה?
iii) השתמשו בשיטה נוספת להוצאת חריגים לבחירתכם (על הדאטה המקורי ולא לאחר סעיף i5.)

6. Predictive model:

i) ממשו שלושה מודלים הנלמד בכיתה והציגו את המדדים הבאים:AUC, recall precision הדאטה לא מאוזן ומומלץ להשתמש במודלים המתאימים למקרה זה
ii) הראו את תהליך מציאת ההיפר פרמטרים במודל, למה בחרתם בפרמטרים אלה?
iii) שימו לב, יש לחלק את הדאטה לtrain & test ביחס של 20%/80%
iv) מומלץ לנסות להשתמש בשיטות להורדת ממדים על מנת לשפר את המודל