מכון טכנולוגי חולון
Holon Institute of Technology

50058 מדע נתונים- תאוריה ומעשה | סמסטר א' תשפ"א

עבודה מספר 1- הערכת שווי רכבים

בעבודה זו נתמקד בבעיית הערכת שווי רכבי יד שניה. לצורך כך נעזר בסט הנתונים אשר פורסם באתר Kaggle-
https://www.kaggle.com/c/1056lab-used-cars-price-prediction/overview.

הנחיות:

ניתוח הנתונים ומיצוי מאפיינים-

א. הורד את סט הנתונים מהאתר (train set) וקרא אותם לתוך data frame.

ב. חלץ את המאפיינים הנומריים שבסט הנתונים. בעבודה זו, נתמקד אך ורק בהם, תוך השמטת המאפיינים שאינם נומריים. שים לב שעבור אותם מאפיינים שבהם מופיעות יחידות, יש להשמיט אותן (לדוגמא באמצעות str.strip).

ג. שרטט גרף של התפלגות הערכים (לכל אחד מהמאפיינים הנומריים) וכן גרף "תרשים קופסה" (box plot). חשב מה אחוז הערכים החסרים.

ד. חשב והצג את הקורלציות ההדדיות בין המאפיינים לבין עצמם ולבין ה- target (המחיר) (ראה דוגמא: https://datatofish.com/correlation-matrix-pandas/).

מדד שגיאה-

ה. בחר מדד שגיאה מתאים עבור הבעיה. הסבר את הבחירה.

פיתוח המודלים-

ו. בחר שניים מבין המודלים שנלמדו בהרצאות. ממש כל אחד מהמודלים והשתמש בו לצורך הערכת שווי רכבים. בחן והשווה את ביצועי המודלים. את הערכת הביצועים כאמור יש לבצע בשלוש שיטות שונות: hold-out, 10-fold cross validation ו- leave-one-out.

הערות:

- יש להגיש דו"ח הכולל פירוט (תרשים + הסבר תמציתי) של הנעשה, תוצאות, ניתוח תוצאות ומסקנות.
- עיקר העבודה הוא ההסברים שלכם- הסבר תמציתי על הנעשה, ניתוח תוצאות ומסקנות.
- מומלץ מאוד לבצע את העבודה באמצעות שפת פייתון.
- התרגיל יוגש ביחידים או בזוגות, באמצעות המודל בלבד. בהגשה בזוגות, מצופה מכל אחד מבני הזוג לשלוט בכל נדבכי העבודה והדו"ח. רק אחד מבני הזוג יגיש את העבודה במודל. יש לרשום שמות + מס' תעודות זהות בראש העבודה.
- עבודות דומות תיפסלנה ויינקטו צעדיים משמעתיים.
- יש להגיש את הדו"ח ב**קובץ PDF אחד בלבד** בהתאם להנחיות.
- עליכם להעלות את הקוד שכתבתם ל- github ולצרף בדו"ח קישור מתאים (אין לצרף את הקוד עצמו לדו"ח). שימו לב שכל המידע וההסברים הרלוונטים צריכים להיכלל בדו"ח (ולא במחברת ה- jupyter או בקוד).
- לוח הזמנים להגשה- בהתאם למוגדר במודול.

מכון טכנולוגי חולון
Holon Institute of Technology

בהצלחה!

צוות הקורס