

## **המחלקה הנדסת תוכנה**

שם הקורס: **סמינר בהנדסת תוכנה**

סמסטר א' תשפ"א

שם הסטודנט אילון מזרחי ת.ז. 260125411

שם הסטודנט דניאל איבקוביץ ת.ז. 316421262

**נושא:** תחרות חיזוי מחירי דירות בארצות הברית.

**קישור לתחרות:**

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

**הפניה למאמר מרכזי:**

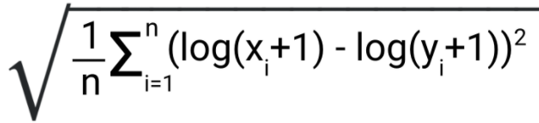
A Hybrid Regression Technique for House Prices Prediction Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, at the Institute of High-Performance Computing (IHPC)

**תיאור הבעיה:**

נתונה קבוצת מאפיינים שמייצגים מידע על דירות המשוכנות במדינת Iowa בעיר Ames שבארה"ב. הבעיה מוגדרת כחיזוי מחירי דירות (בעיית רגרסיה) של קבוצת רשומות מבחן, על סמך קבוצת האימון הנתונה. לבעיה זו קשיים והיביטים שונים:

1. לא כל המאפיינים הנתונים תורמים בהכרח לתוצאת החיזוי על קבוצת המבחן, והם עלולים אף לגרוע מהתוצאה.
2. הנתונים אינם בפורמט מתאים בהכרח לאלגוריתמי רגרסיה הקיימים.
3. הנתונים אינם מנורמלים וטווח הערכים של כל מאפיין משתנה.
4. ישנם מגוון סוגי אלגוריתמים לביצוע רגרסיה, איזה מהם הוא\הם המתאימ\ים ביותר לנתוני הבעיה?
5. מספר המאפיינים קטן יחסית למספר הרשומות הנתונות.

פתרון הבעיה נמדד לפי ערך ה- RMSLE, שורש ממוצע סכום ריבועי השגיאות בין ערכי הלוגריתמים של הערך החזוי לערך האמת, לפי הנוסחה:



Where xi is the i'th prediction of the hypothesis on the given data, and yi is the ground truth.

**מטרת הפרויקט:**

מטרת הפרויקט הינה לחקור שיטות שונות בתהליך רגרסיה, החל מתהליך הכנת הנתונים לאימון ועד לחקר המודלים השונים לביצוע חיזוי של ערכי מאפיין המטרה של הבעיה: מחירי דירות.

**תיאור הנתונים:**

קיימים 79 מאפיינים שמתארים 1460 דירות (מספר הרשומות) במדינת Iowa, בעיר Ames שבארה"ב. נתון בנוסף משתנה המטרה "מחירי הדירות". לפי מטרת הפרויקט, באמצעות הנתונים כקלט, יש לייצג בצורה כלשהי את משתנה המטרה. ייצוג זה יהווה את ההיפותזה האופטימלית (אותה יש לחפש) שממפה נתוני דירה לפי 79 המאפיינים אל המחיר שלה.

**מתודולוגיה:**

ניתן לחלק את השיטות בהן השתמשנו על מנת לפתור את הבעיה לכמה שלבים:

1. הכנת הנתונים:
   1. הצגת התפלגויות ערכי כל המאפיינים על מנת להתמיר את הנתונים לפי הצורך אל ערכים שמתאימים לאלגוריתמים שונים לביצוע רגרסיה.
   2. חקירת outliers עבור כל מאפיין: הצגת ערכי המאפיינים אל מול ערכי משתנה המטרה (ניסיונות חיזוי איתם ולבעדיהם).
   3. למידת משמעויות המאפיינים (בנוסף, ציור מפת קורלציית כל המאפיינים עם משתנה המטרה) על מנת להוסיף\להחסיר\להרכיב מאפיינים מסויימים לפי הצורך.
   4. בדיקה והוספה של ערכים חסרים עבור כל המאפיינים. הוספה של ערכים "הגיוניים" (לפי ממוצע, היגיון וכו'...).
   5. המרת פורמטים למאפיינים מסויימים לפי הצורך (מאפיין נומרי – מאפיין קטגורי ולהיפך).
2. הכנת מודלים:
   1. חקירת מודלים שונים של למידת מכונה לביצוע רגרסיה. המודלים התחלקו לסוגים שונים:
      1. מודלים עבור רגרסיה לינארית:

Lasso, Ridge, ElasticNet, Kernel Ridge Regression (a ridge linear model that uses the

Polynomial Kernel trick to separate the data in a higher-dimensional space).

* + 1. מודלים שאינם בהכרח עבור בעיית רגרסיה לינארית:

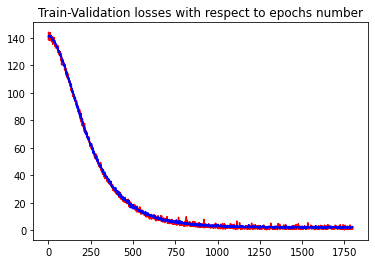
Gradient Boosting algorithms (XGB, LGB), a deep neural network (using MSE criterion).

* 1. בחירת היפר-פרמטרים: בחירת פרמטרים לפי המאמר עליו הסתמך הפרויקט, ניסוי וטעייה, מדידת ציון שגיאת ה- RMSLE על מנת למדוד הצלחת כל מודל על הולידציה, הגשות חוזרות על מנת לקבל ציונים שונים בתחרות על ה- test set המלא, שימוש באלגוריתם GridSearchCV על מנת לחפש ערך אופטימלי עבור פרמטר ה- learning-rate.
  2. הרכבת מודלים שונים (ensemble) ובחינת המודלים המורכבים לפי המטריקות שהוצגו בסעיף 2.2. ההרכבה נעשתה בהשראת המאמר המוצג, ובהשראת שתי שיטות:
     1. Model Averaging (ממוצע משוקלל ולא משוקלל- ערך זהה למשקולות).
     2. שיקלול המודלים השונים לפי החלטת מודל נוסף (meta-model) אשר חוזה את משקולותיהם. ה- meta model התאמן על סמך שגיאותיהם של כל המודלים על הולידציה באמצעות חלוקת K-Fold Cross Validation, ל- 5 folds.
     3. ניסוי וחקירה של network pruning עבור מודל רשת הנוירונים, במטרה לאזן את שגיאת variance (אם יש צורך) כיוון שמודל זה עלול להיות רגיש לשגיאה זו.

לאחר סקירת המתודולוגיות השונות יחד עם התוצאות השונות, הוחלט להשתמש באלגוריתם X-Gradient Boosting בלבד עבור ההגשה הסופית, כיוון שהניב את התוצאה הטובה ביותר.

**תוצאות:**

1. תוצאות שגיאת האימון של רשת הנוירונים על נתוני האימון ועל נתוני הוולידציה לפי מספר האיטרציות (epochs):

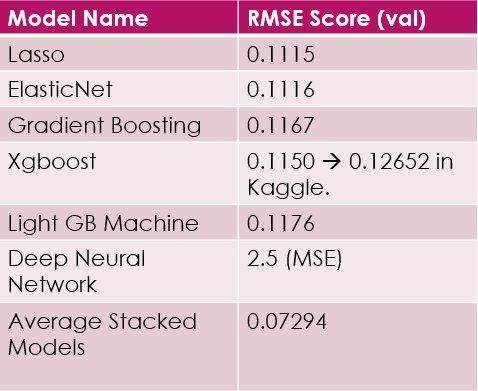


1. תוצאות שגיאת מודל ה- XGB על קבוצות האימון והוולידציה, לפי ערך משתנה של גדלי הקבוצות. לבסוף תוצאות כל האימונים עוברות מיצוע לערכים בטווח [0, 1].



1. תוצאות השגיאות על קבוצת הוולידציה עבור כל אחד מהמודלים ועבור מודל ה- ensemble שכלל בתוכו:

ElasticNet, Gradient Boosting Regressor, Lasso as base models and Kernel Ridge Regressor as the meta-model.

****

1. התוצאה הטובה ביותר שהשגנו בתחרות, על ידי מודל ה- XGB:

Rank: 1396/5970 participants Score: 0.12652

**רשימת מקורות:**

1. The main paper:

A Hybrid Regression Technique for House Prices Prediction. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh at the Institute of High-Performance Computing (IHPC).

1. Other references:
   1. Complete Guide to Parameter Tuning in XGBoost with codes in Python. By Aarshay Jain, MS in Data Science at Columbia University.
   2. House Prices Prediction Using Deep Learning. By Data Scientist Masha Mir.