מגיש: ניר דעוס 315782987

# מטלה 2 - מודל חיזוי מחירים

מטרת המטלה היא לבצע ניתוח נתונים באמצעות הכלים שלמדנו בשיעור על מנת לחזות מחירי הרינג.  
הנתונים הם טבלה של 1000 תצפיות של מאפייני לקוחות חנות דג הרינג.

שלב א – EDA

כשלב מקדים לחיזוי אנו נדרשים לעבד את המידע.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature1** | **Feature2** | **Feature3** | **Feature4** | **Feature5** | **Feature6** | **Feature7** | **Feature8** | **Feature9** | **Feature10** | **Price** |
| 3.84 | 2.63 | A | Apartment | 3.67 | 3.51 | 2.50 | 2.14 | 2.65 | 3.52 | **604.63** |
| 4.06 | 3.73 | B | Apartment | 4.31 | 5.43 | 2.23 | 2.23 | NaN | 2.77 | **965.57** |
| 2.56 | 4.08 | C | House | 3.51 | 5.41 | 2.47 | 2.54 | 2.46 | 2.47 | **623.97** |
| 4.08 | 5.55 | B | House | NaN | 6.00 | 2.24 | 3.91 | 3.72 | 2.56 | **2270.05** |
| 4.80 | 4.47 | A | Apartment | 5.65 | 4.20 | 3.01 | 2.31 | 2.91 | NaN | **1958.46** |

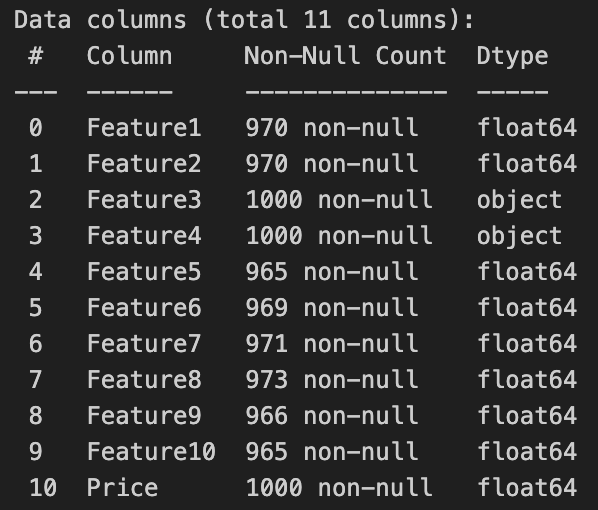
ראשית, הסתכלות פשוטה בשורות הראשונות:

כבר מהסתכלות ראשונית ניתן להבחין שיש שלנו 10 מאפיינים (שדות). מתוכם, 8 מאפיינים נומריים ו-2 מאפיינים קטגוריאליים. בנוסף ניתן להבחין שיש ערכים חסרים במאפיינים 5,9,10 ויתכן שיש גם במאפיינים נוספים.

כדי לגלות ליקויים נוספים, נשתמש בפונ׳ describe שתוציא דוח מסכם על הסטטיסטיקות התיאוריות של המאפיינים הנומריים:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Feature1** | 970 | 5.30 | 31.99 | 1.20 | 3.59 | 4.25 | 4.92 | 1000.00 |
| **Feature2** | 970 | 3.52 | 32.27 | -1000.00 | 3.87 | 4.56 | 5.21 | 7.74 |
| **Feature5** | 965 | 5.22 | 32.07 | 1.29 | 3.48 | 4.21 | 4.86 | 1000.00 |
| **Feature6** | 969 | 4.66 | 1.01 | 1.51 | 3.94 | 4.63 | 5.36 | 7.50 |
| **Feature7** | 971 | 2.76 | 0.59 | 2.00 | 2.30 | 2.64 | 3.09 | 5.24 |
| **Feature8** | 973 | 1.78 | 32.15 | -1000.00 | 2.33 | 2.68 | 3.17 | 5.85 |
| **Feature9** | 966 | 2.80 | 0.61 | 2.00 | 2.32 | 2.65 | 3.18 | 5.24 |
| **Feature10** | 965 | 2.81 | 0.60 | 2.00 | 2.33 | 2.69 | 3.19 | 5.93 |

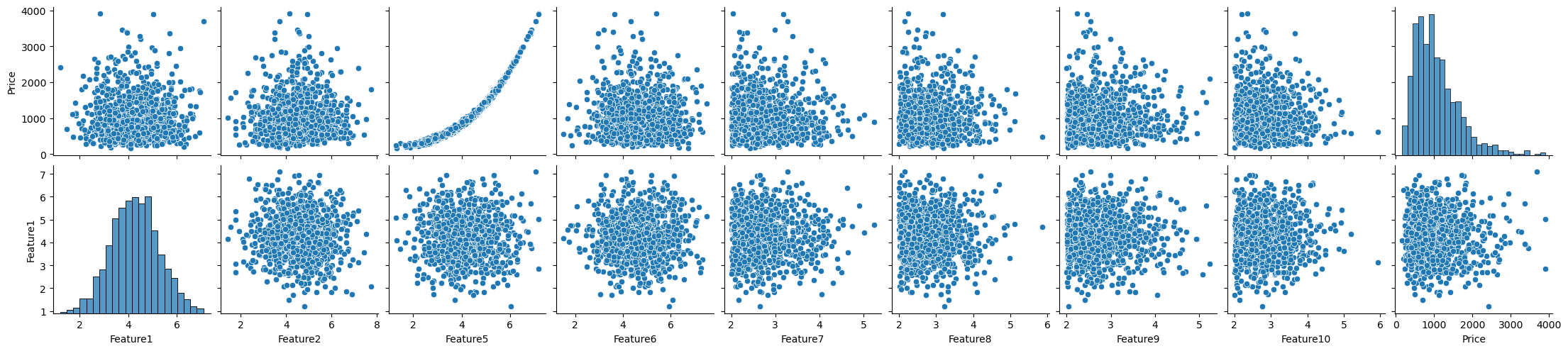
ניתן להבחין שיש שערכים חברים בכל שדה (הצפייה היא ל-1000 ערכים) ושיש ערכי קיצון שמשפיעים על שאר הסטטיסטיקות. הערכים הם כה קיצוניים כך שניתן להניח שמדובר בטעות מבלי שניתן לשחזר את הערכים המקוריים במקומם.

כדי לראות האם יש ערכים חסרים גם במאפיינים הקטגוריאליים נשמש בפונ׳ info שנותנת גם מידע על סוגי הנתונים (data types).

הערכים הקטגוריאלים (שדות 3,4) מכילים כל אחד 1000 תצפיות ולכן תקינים מבחינה זו.

כעת נדרש לחשוב איך צריך להשלים את הערכים האחרים מכל אחד מהשדות.

נשרטט גרף התאמה בין כל שדה למחיר:

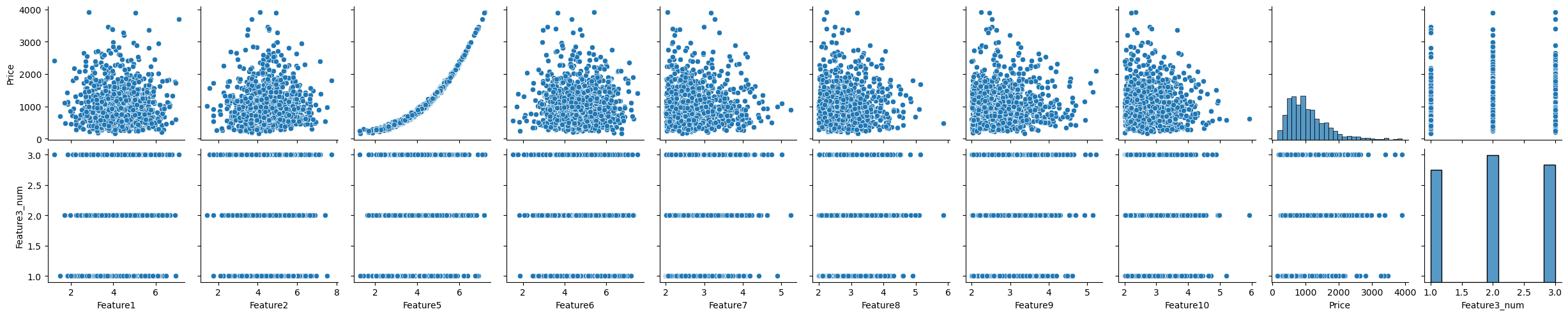


גרף זה מראה בצורת ההתאמה בין כל שדה למחיר באמצעות דיאגרמת פיזור. ניתן להבחין מכאן ש:

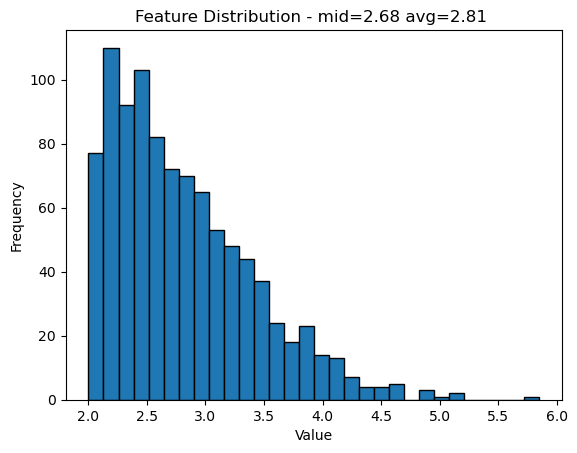
1. שדה 5 הוא בעל קשר פולינומי למחיר. אם נמצא פונקצייה רגרסיה מקורבת נוכל להשלים את הערכים החסרים בשדה זה במידה התאמה גבוהה לערכים המקוריים.
2. שאר השדות (הנומריים) אינם בעלי התאמה גבוהה למחיר, ובאופן פשטני יהיה נכון להניח השלמת ערכים חסרים לפי ממוצע או חציון.
3. לא מסתמן קשר לינארי או אחר בין שדה אחד לשדה אחר שאינו המחיר. (לא הבאתי כאן את הגרף המלא)

בעקבות הקשר הברור בין שדה 5 למחיר, אימנתי מודל רגרסיה פולינומית כאשר נמצא ש- degree=4 נותן את ערך השגיאה המוחלטת הנמוך ביותר (mae=0.07). באמצעות המודל ביצעתי תחזית והשלמתי את הנקודות החסרות משדה זה.

בדקתי לאחר מכן אם יש קשר בין השדה הקטגוריאלי לבין אחד השדות הנומריים ע״י המרת הערכים שלו למספרים - {'A' : 1, 'B' : 2, 'C' : 3}:



ניתן לראות שאין קשר – אחרת היינו רואים גרף מדרגות.

כעת ניתן להשלים את הערכים החברים באמצעות החציון או הממוצע. שרטטתי בהיסטוגרמה את התפלגות שלהם כדי להבין האם הממוצע קרוב לחציון ובאיזה מרחק. לדוגמא שדה 8:

כאן בחרתי להשלים את הערכים החסרים לפי החציון כיוון הוא קרוב יותר למרבית הערכים מאשר הממוצע.

לאחר השלמת כל הערכים, נדרש להמיר את השדות הקטגוריאליים לערכים נומריים. נמיר אותם לערכים בינאריים 1/0 באמצעות פונקציית get\_dummies. כיוון שהם ערכים משלימים (אם זה 0 השני 1) ניתן להוריד לגמרי עמודה אחת מכל סוג בשדה קטגוריאלי.

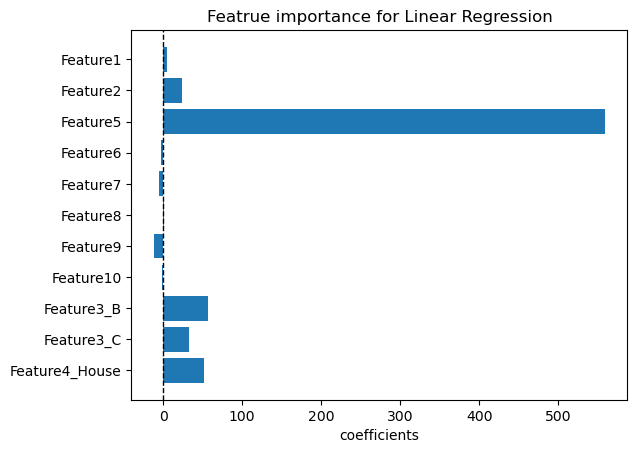
pd.get\_dummies(train\_data, columns=["Feature3", 'Feature4'], drop\_first=True,dtype = int)

שלבים ב – ה: בחינה ואימון מודלים לתחזית המחיר

1. רגרסיה לינארית

בחנתי מודל של רגרסיה לינארית על הנתונים באמצעות k-fold. ער השגיאה שהתקבל:  
mae = 124.33.

כדי להבין איזה שדות השפיעו על התחזית נשרטט את ה-feature importance של המודל:



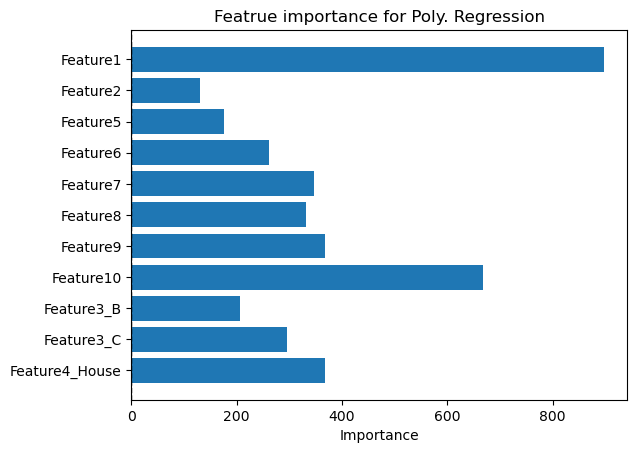
כצפוי, שדה 5 השפיע בצורה הכי ניכרת על המחיר. שדות 6-10 השפיעו בצורה שלילית על פונ׳ המטרה.

1. רגרסיה פולינומית

כדי לקבוע את המעלה של הרגרסיה (חזקה) הרצתי לולאה על הערכים 2-10 עד מציאת השגיאה המינימלית מביניהם. התוצאות:

\*\* best d=3, mae=8.51 \*\*

כדי להבין איך המודל עבד, סכמתי את ערכי המקדמים בערך מוחלט של כל שדה ושמתי בגרף:



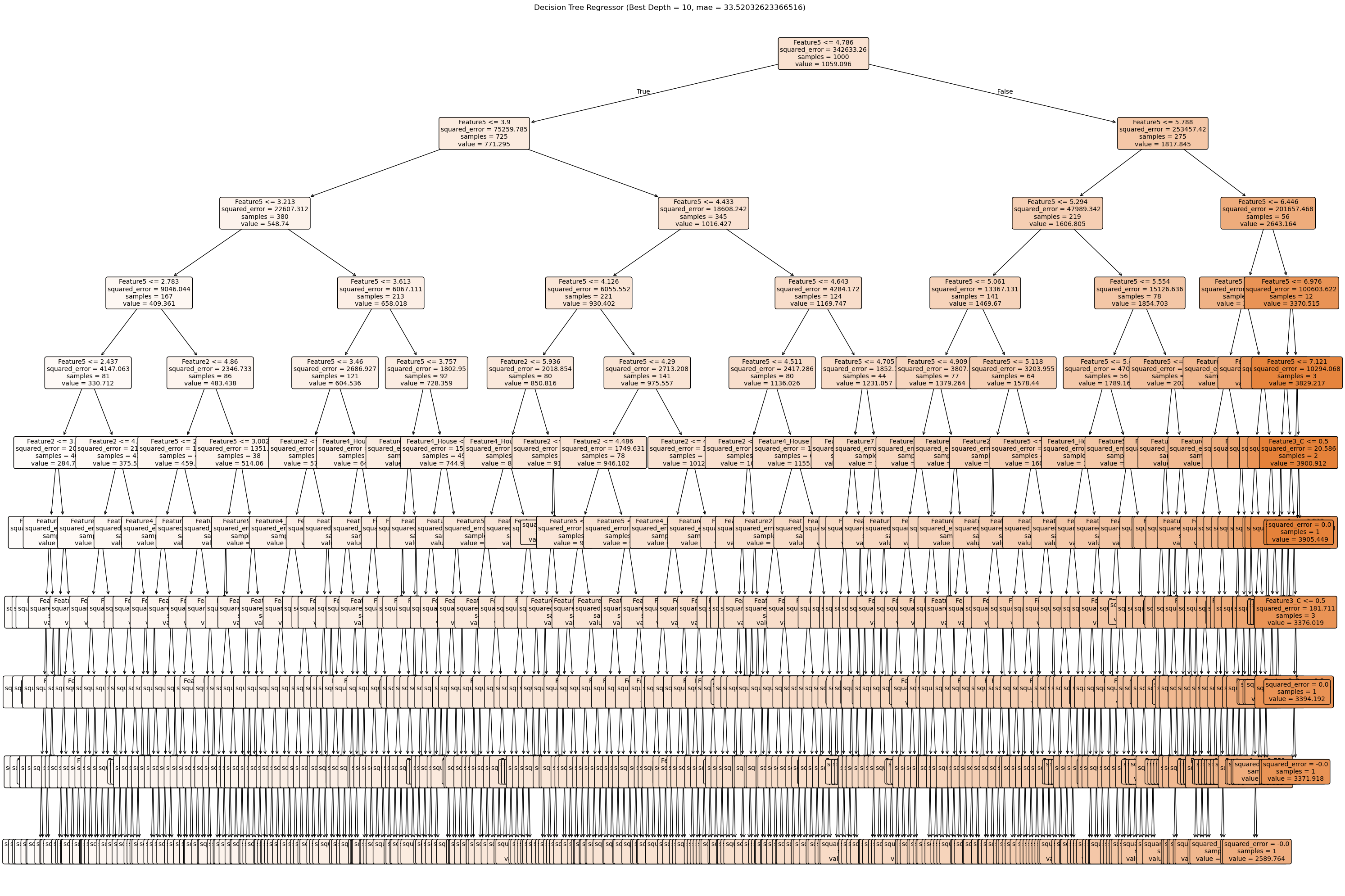
הפעם, שדה 1 השפיע הכי הרבה על פונקציית המטרה ושדה 5 זניח.

1. עץ החלטה

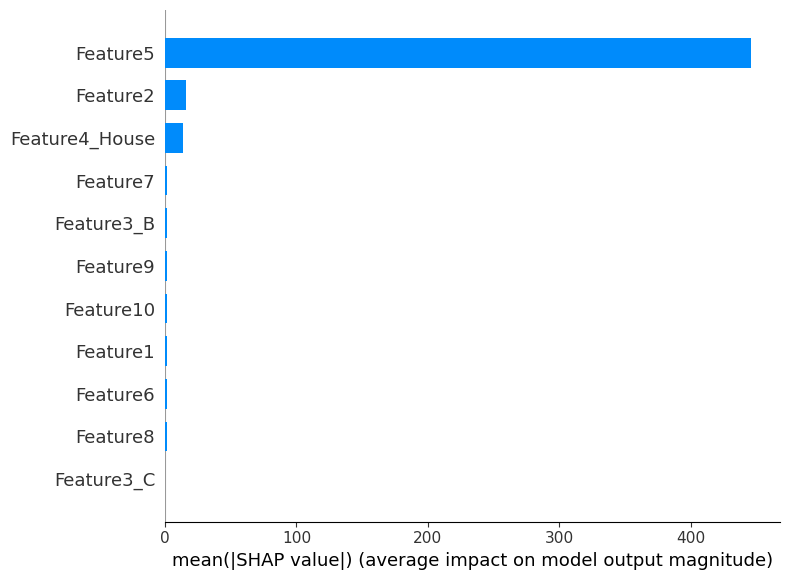
כדי לקבוע עומק אופטימלי של העץ הרצתי לולאה על הערכים 2-30. תוצאה סופית:

\*\* best depth=10, mae=33.52 \*\*

שרטוט של העץ:



כדי להבין איך המודל עובד, שרטטתי את ה-feature importance באמצעות שיטת shap:



גם כאן שדה 5 המשפיע ביותר.

1. Random Forest

כדי לאמן מודל כזה צריך להחליט מספר פרמטרים, החשובים: עומק העת ומספר העצים. באופן פשוט (אך לא יעיל) ניתן לבצע לולאה על הערכים:

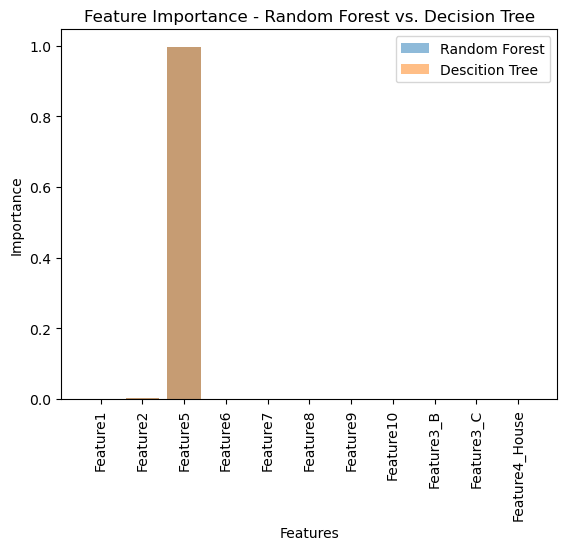
עומק – לולאה בין 2 ל-20 בקפיצות של 2.

מס׳ עצים - לולאה בין 100 ל-1000 בקפיצות של 200.

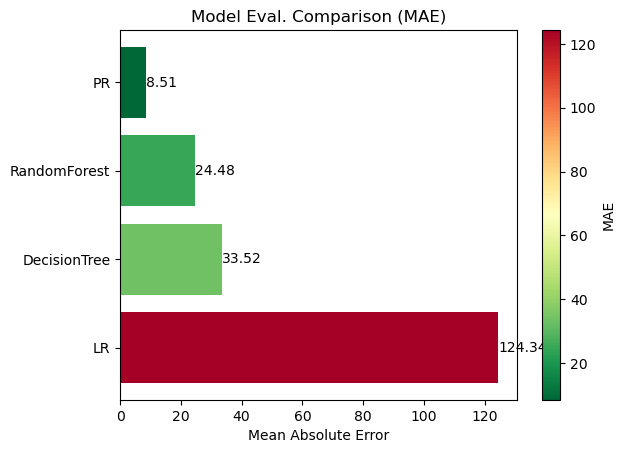
שיטה זו הניבה את התוצאות הבאות:

\*\* best depth=12, best\_n\_estimators=300, mae=24.48 \*\*

שירטטתי את ה-feature importance בהשוואה לעץ ההחלטה:

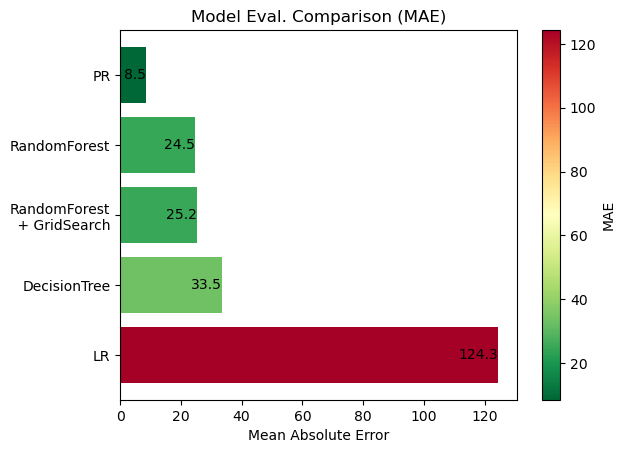


לצורך השוואה מסכמת ולראת השלה הבא שרטטתי ביחד את ה-mae שיצא מכל מודל:



ניתן לראות שהמודל הפולינומי מוביל עם שגיאה ממוצעת של 8.5.

רציתי לשפר מעט את המודל RandomForest באמצעות GridSearch אך התוצאות היו זניחות:



שלב ו׳ – מסקנה

למסקנה – קיבלנו **שהרגרסיה הפולינומית במעלה 3** מביאה את השגיאה הכי נמוכה (8.5) בהשוואה לכל מודל אחר שאימנו.