RTB Challenge-StartApp, 18.9.19

כללי:

בהמשך לתהליך ראיון העבודה שאני נמצא בחברתכם, התבקשתי לבצע את מטלת הבית בתחום RTB. להלן עיקרי הממצאים מהמטלה.

הבנת הנתונים:

1. בחינת והתאמת המשתנים הבלתי תלויים על המשתנה התלוי- הבנה עסקית
2. רוב המשתנים קטגוריאליים-חשיבה על מודל מותאם לכך (RF)

עיבוד הנתונים:

1. הסרת רשומות בעלי ערכים חסרים, סה"כ רשומות לאחר ההסרה: 200K.
2. ביצוע התמרה למשתנה Clicks
3. ביצוע התמרה למשתנה impressions
4. הגדרה מחדש של המשתנה edate
5. הגדרת המשתנה product כפקטור
6. התמודדות עם משתנים קטגוריאלים בעלי ערכים מרובים- חלוקה מחדש של מספר הערכים בכל משתנה מרובה ערכים.

מידול:

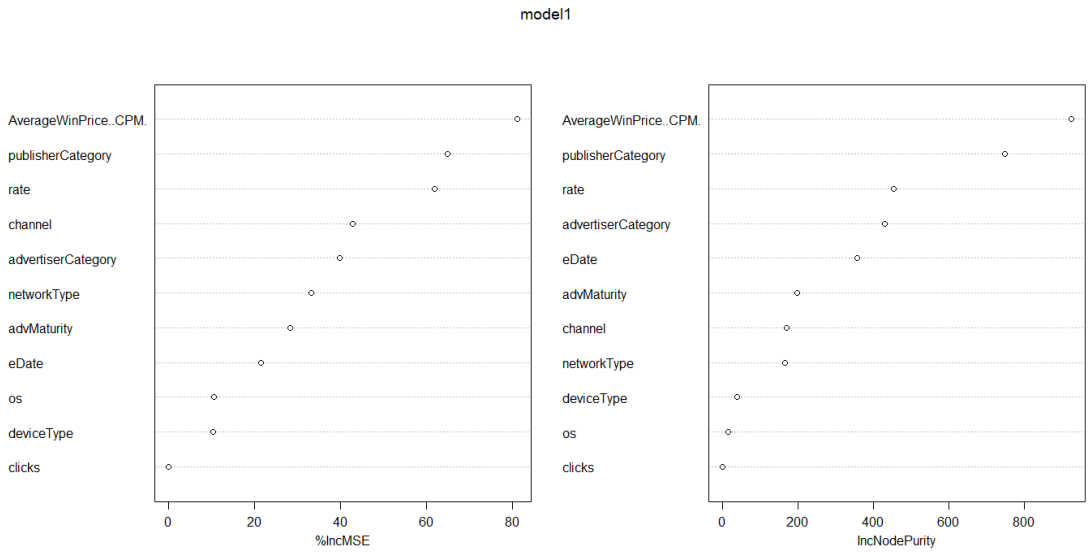
1. בחירת המשתנים המסבירים למודל
2. התמודדות עם בעיית זיכרון במחשב
3. חיפוש אחר הפרמטרים הנכונים שיסייעו לחיזוי טוב יותר על בסיס הנתונים הקיימים.
4. הרצת המודל עם הפרמטרים המותאמים

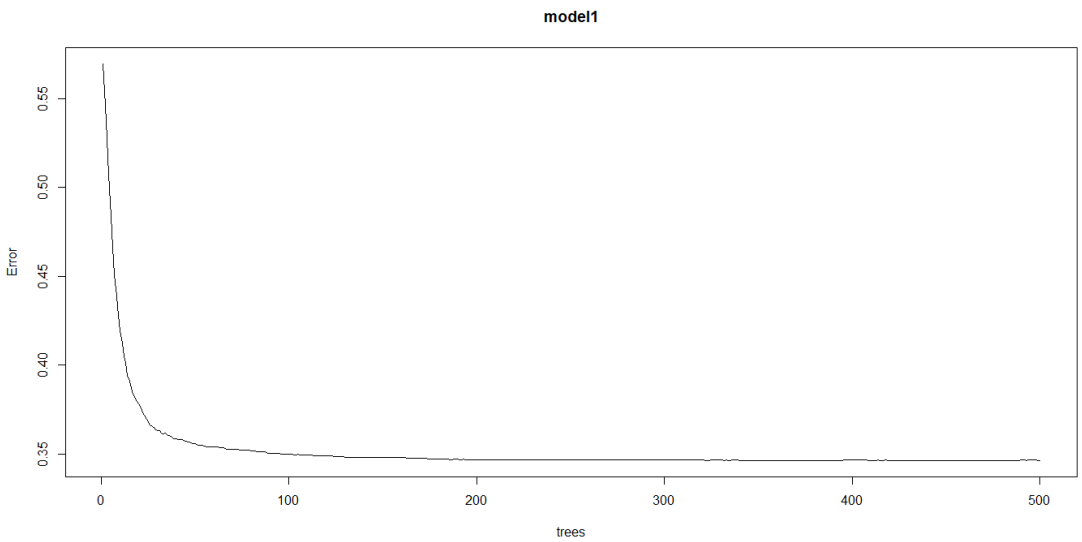
תהליך ותוצאות:

1. בחרתי להריץ את המודל עבור 20000 רשומות ולא על כל הנתונים, משיקולי זמן ריצה. הנתונים שנבחרו, נבחרו באופן אקראי.
2. תחילה נבחן מודל ללא משתנים קטגוריאליים מרובי ערכים. כמו כן נבחן מודל RF בסיסי ללא אופטימיזציה, (מודל 1). תוצאות המודל מובאות להלן:

```
model1 <- randomForest(Log_impressions ~ eDate+channel+os+networkTyp
e+deviceType+publisherCategory+advertiserCategory+advMaturity+rate+c
licks+AverageWinPrice..CPM., data = Train, importance = TRUE)
> model1

Call:
 randomForest(formula = Log_impressions ~ eDate + channel + os +
networkType + deviceType + publisherCategory + advertiserCategory +
```

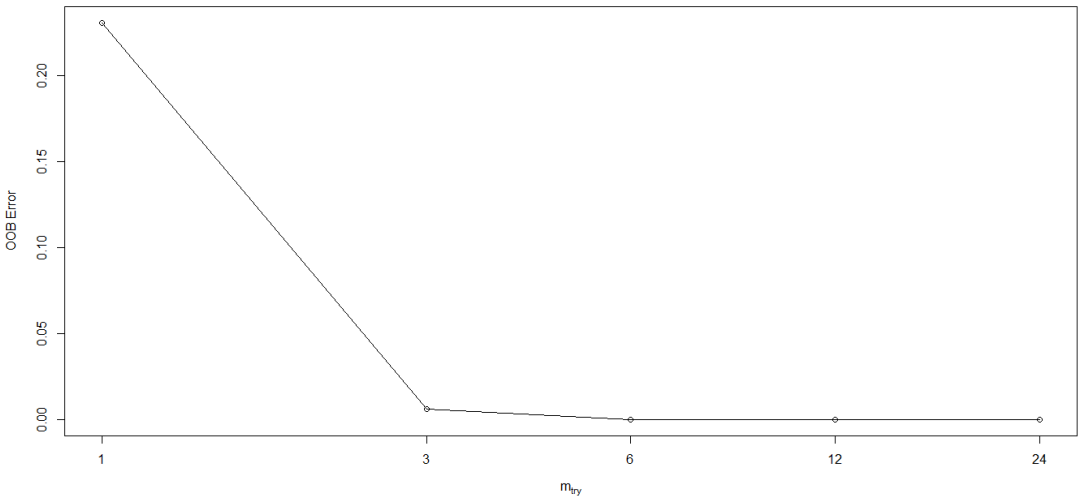
```
advMaturity + rate + clicks + AverageWinPrice..CPM., data = Train,
importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 0.3465639
                    % Var explained: 14.95
> plot(model1)
> importance(model1)
                      %IncMSE IncNodePurity
eDate                21.39620     357.05088
channel              42.85181     169.37809
os                   10.46205      13.69771
networkType          33.18383     163.84758
deviceType           10.38517      37.12939
publisherCategory    64.92110     747.78515
advertiserCategory   39.79895     430.21011
advMaturity          28.27980     197.17707
rate                 61.85657     452.59078
clicks                0.00000       0.00000
AverageWinPrice..CPM. 81.22061     925.01888
> varImpPlot(model1)
```

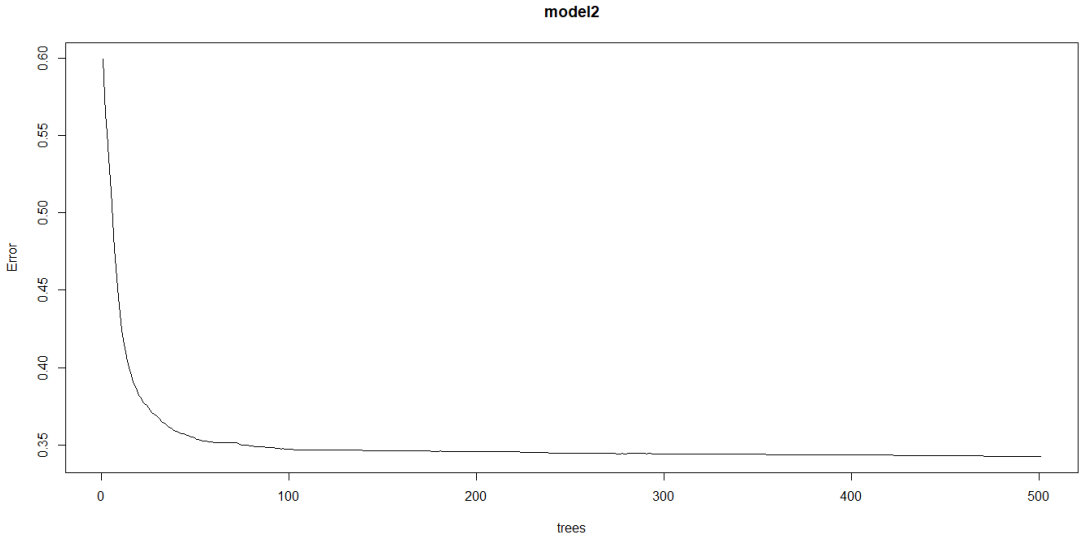




בוצע כיוונון של הפרמטרים עבור RF על ידי שימוש בכלי אופטיזציה. הערכים שהתקבלו כאופטימליים ביותר הינם Mtry=6, ntree=400. ניתן לראות את הגרפים לבחינת ערכים אלו ולבחון את השיפור בהתאם.

```
> # Fine tuning parameters of Random Forest model
> tune.rf <- tuneRF(Train[,-c(1,5,11:14,16,18)],Train[,19], stepF
actor=0.5)
mtry = 3  OOB error = 0.006381028
Searching left ...
mtry = 6    OOB error = 0.0003839671
0.9398268 0.05
mtry = 12   OOB error = 0.0001517419
0.6048051 0.05
mtry = 24   OOB error = 0.0002144148
-0.4130235 0.05
Searching right ...
mtry = 1    OOB error = 0.2307623
-1519.756 0.05
```



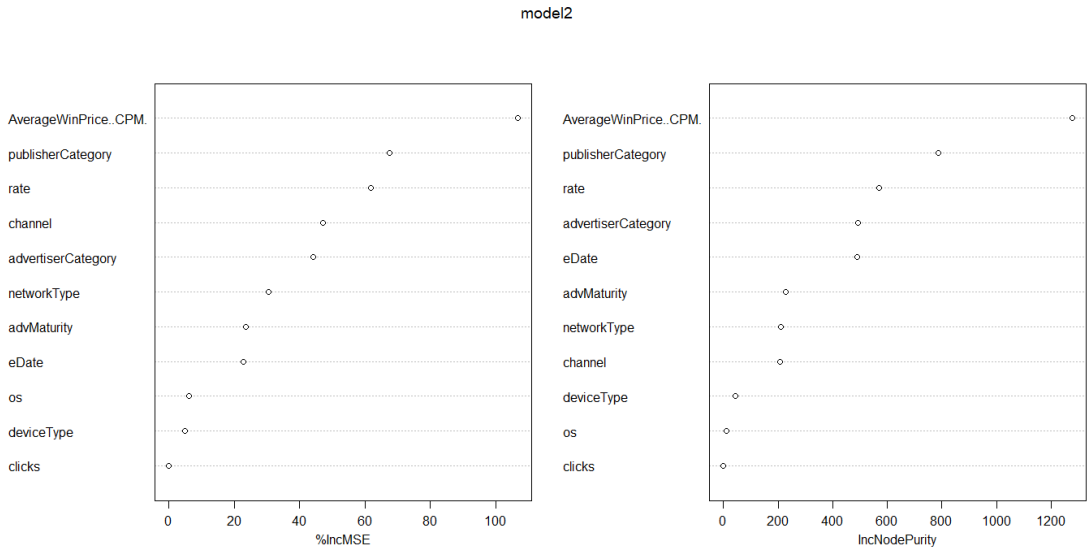


3. הרצתי את המודל שנית עם אותם משתנים מסבירים שהרצתי במודל 1. מודל זה נקרא מודל 2.

```
model2

Call:
 randomForest(formula = Log_impressions ~ eDate + channel + os +
networkType + deviceType + publisherCategory + advertiserCategory
+      advMaturity + rate + clicks + AverageWinPrice..CPM., data
= Train,      ntree = 501, mtry = 6, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 501
No. of variables tried at each split: 6

          Mean of squared residuals: 0.3416829
                    % Var explained: 16.15
```

4. עבור כל מודל, בחנתי אילו משתנים השפיעו בצורה הטובה ביותר. בגרף הבא מוצגים המשתנים המסבירים על פי סדר ההשפעה שלהם:

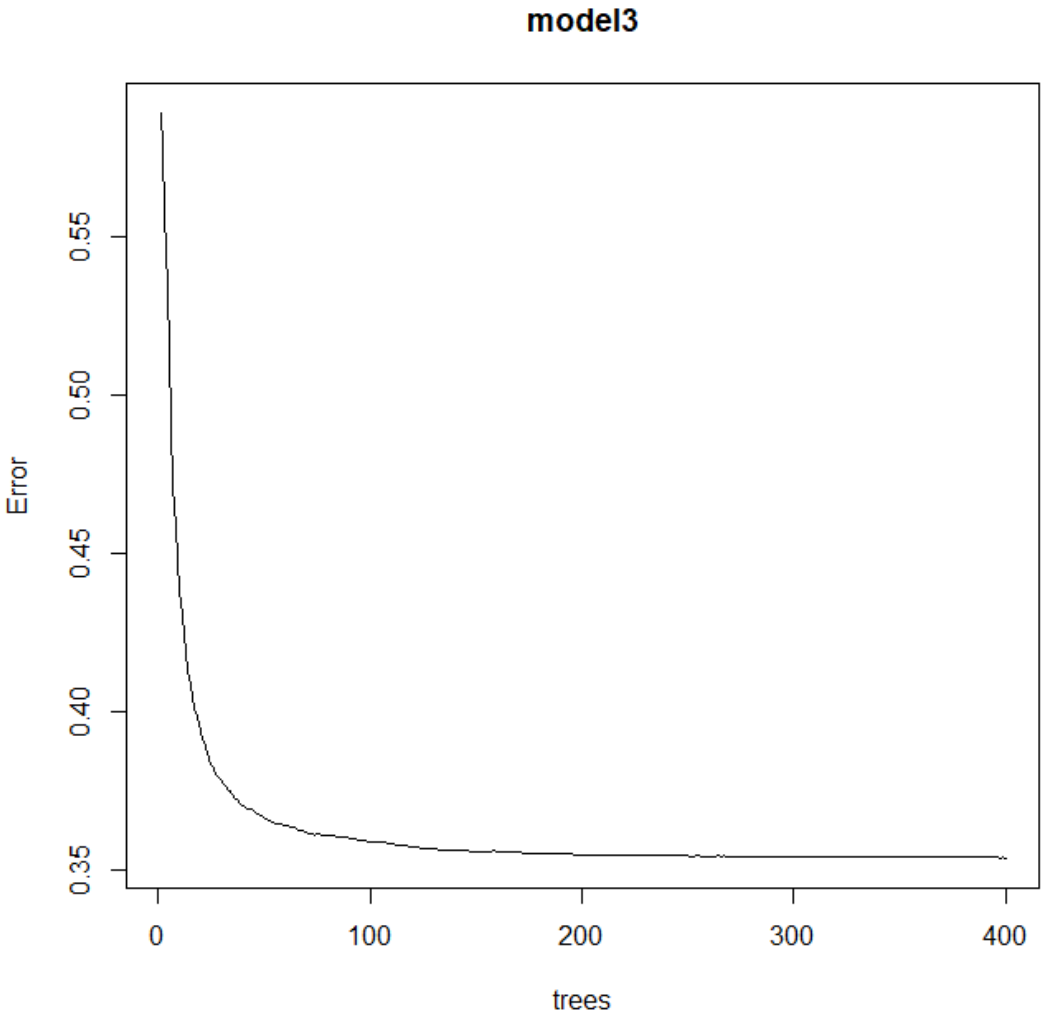
```
importance(model2)
                      %IncMSE IncNodePurity
eDate                24.242340    489.374506
channel              48.276513    205.767371
os                    5.195240      9.191813
networkType          30.951983    209.751535
deviceType            3.752109     41.581697
publisherCategory    71.900502    787.914841
advertiserCategory   44.232490    493.999130
advMaturity          23.788185    227.720271
rate                 65.344918    571.017826
clicks                0.000000      0.000000
AverageWinPrice..CPM. 99.886687   1274.863688
> varImpPlot(model2)
```

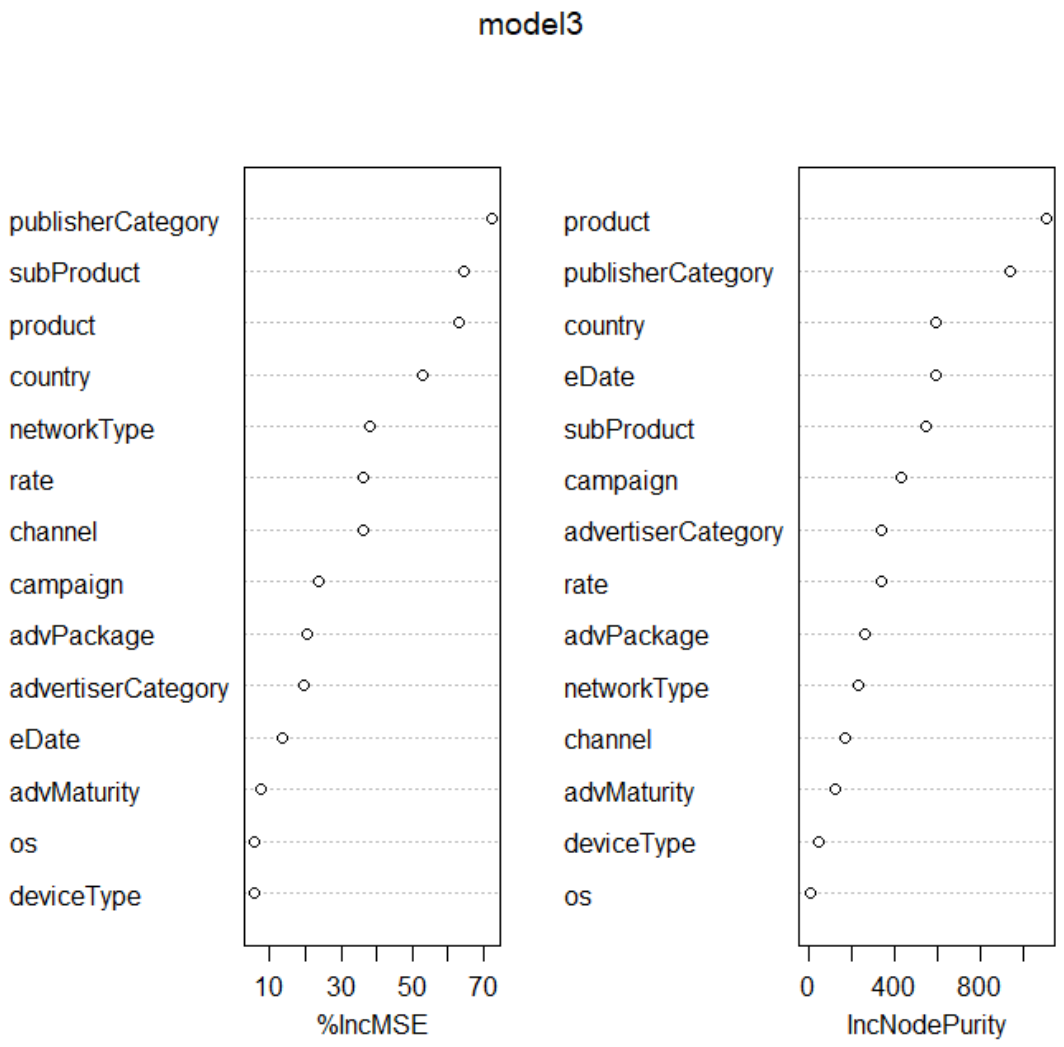


5. בחרתי לבצע חלוקה מחדש של המשתנים הקטגוריאלים בעלי ערכים מרובים. לאחר מכן הכנסתי אותם למודל (מודל 3) והרצתי שוב.

6. להלן פירוט המודל ותוצאותיו:

```
> model3 <- randomForest(Log_impressions ~., data = Train, ntree
= 400, mtry = 6, importance = TRUE)
> model3

Call:
 randomForest(formula = Log_impressions ~ ., data = Train, ntree
= 400,      mtry = 6, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 400
No. of variables tried at each split: 6

          Mean of squared residuals: 0.3540045
                    % Var explained: 13.72
```


model3
Error
0.35
0.40
0.45
0.50
0.55
0
100
200
300
400
trees



מהגרפים ניתן ללמוד על חשיבות המשתנים ועל השפעת הפרמטר ntree על המודל.

7. בשלב הבא, המטרה להריץ את נתוני האמת על המודל שנבנה. השגיאה שהתקבלה עבור חלק מהרשומות היא שישנם נתונים בסט הבחינה שלא קיימים בסט האימון. ישנם מספר פתרונות לכך, ייתכן והרצה מלאה של כל נתוני האימון יפתרו את הבעיה וזאת מכיוון שהרצתי רק על 20 אלף נתונים (10% מהנתונים המלאים). במידה וגם זה לא יפתור את הבעיה ניתן ליצור פתרונות אחרים כגון השתלה של כל הערכים בסט האימון.

בברכה,

לירן בן ציון

052-3068352