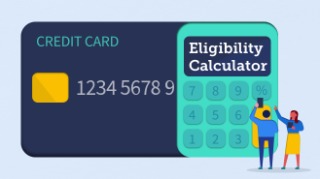
****

המחלקה להנדסת תעשייה וניהול

Machine Learning Project

תמונה שמכילה ציור

התיאור נוצר באופן אוטומטי

מנחה הקורס: ד"ר רוני הורביץ

שמות חברי הצוות:

אוראל מזרחי

יאיר בר לב

אופיר לוין

עדיאל רחמים

קישור ל- bitbucket : <https://bitbucket.org/ofir_levin1313/final-project/src/master/R%20PROJECT>

**מבוא**

בפרויקט זה, במסגרת הקורס "כריית ידע ולמידת מכונה", נדרשנו למצוא מסד נתונים המכיל לפחות 10 אלף רשומות ממידע אמיתי (לא מסד לימודי). בחיפוש בגוגל מצאנו מסד נתונים על לווים באשראי ממדינת טאיוון. המסד נועד לטובת מודלים לניהול סיכונים בתחום אי החזרת הלוואה על ידי הלקוחות ב**חודש הקרוב** בטאיוואן, הנתונים הם לשנת 2005.

**מטרת המודל** – זיהוי לקוחות אשר לא יעמדו במסגרת האשראי שלהם בחודש הקרוב.

מטרת הפרויקט – יצירת מודל כלכלי אשר יוכל לצפות את הלווים שלא יעמדו במסגרת האשראי שלהם.

**מבנה מסד הנתונים:**

מספר רשומות: 30K , מספר תכונות: 24.

|  |  |
| --- | --- |
| **שם** | **הסבר** |
| default payment next month | יעמוד בהחזר התשלום בחודש הקרוב( עמודה בינארית (כן\לא - 1\0).(עמודת המטרה). |
| Limit\_bal | סכום האשראי אשר ניתן לו ולמשפחתו |
| SEX | מין |
| Education | רמת השכלה (1-ביתספר 2-אוניברסיטה 3 תיכון 4 אחרים) |
| Marrige | מצב משפחתי (1- נשוי 2- רווק 3- אחרים) |
| Age | גיל (בשנים) |
| Pay\_0 – pay\_6 | מראות את מצב התשלום בחודשים הקודמים מאפריל 2005 עד ספטמבר 2005 (0= שילם כהלכה 1 = איחר בחודש, 2 = איחר חודשיים ......9= איחור של 9 חודשים ויותר) |
| Bill\_amt1 - Bill\_amt6 | סכום הצהרת חוב בחודשים אפריל עד ספטמבר בהתאמה. |
| Pay\_amt1 - Pay\_amt6 | סכום תשלום הקודם עבור החודשים אפריל עד ספטמבר בהתאמה. |

**הסבר על העמודות:**

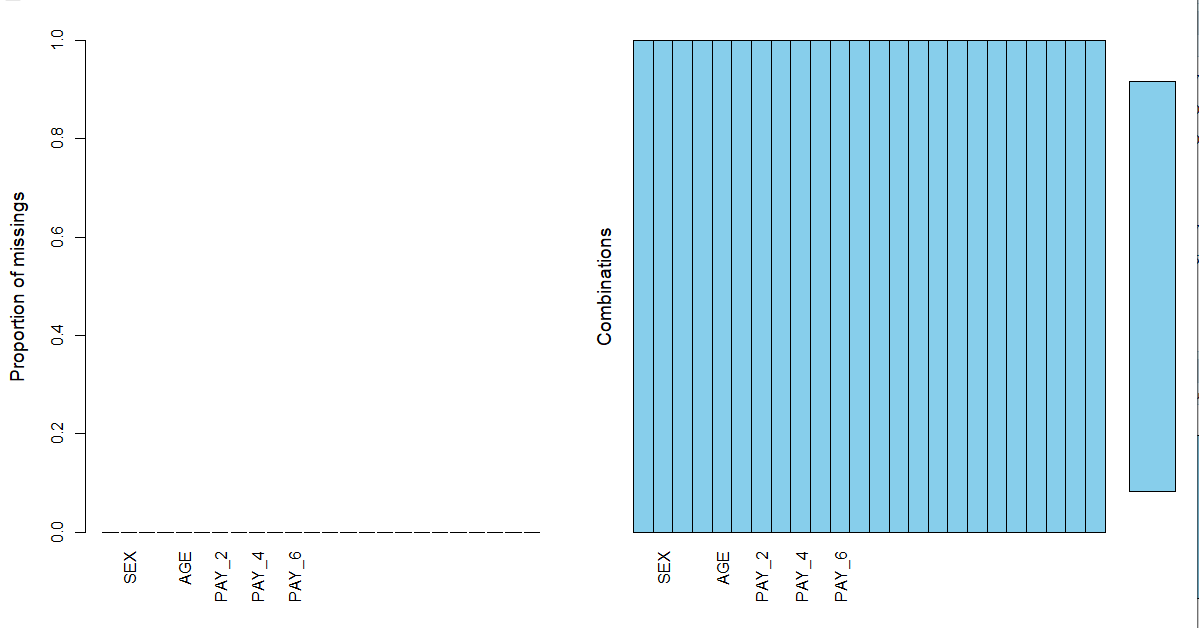
**מטרות הפרויקט**

* הרחבת הידע בחומר הנלמד, תוך שימוש בלמידת מודלים ונושאים חדשים.
* הפקת מודלים אשר באמצעותם יתקבלו תוצאות חיזוי אופטימאליות שבאמצעותן נצליח לחזות את עתידה הכלכלי של החברה.
* זיהוי עמודות (תכונות ) אשר משפיעות בצורה מובהקת על עמודת המטרה.
* מתן כלי חישובי עבור מקבלי החלטות על מנת לצמצם את נתינת האשראי ללקוחות אשר צפויים לא להחזיר.

**תהליך העבודה – EDA**

**ניקיון הנתונים** – בשלב הראשון נרצה לבחון את שלמות הנתונים ולהפוך אותם לסוג המתאים במידת הצורך (FACTOR NUMERIC וכו').

תחילה נבדוק האם יש ריקים על ידי שימוש בפונקציה **aggr(df)**

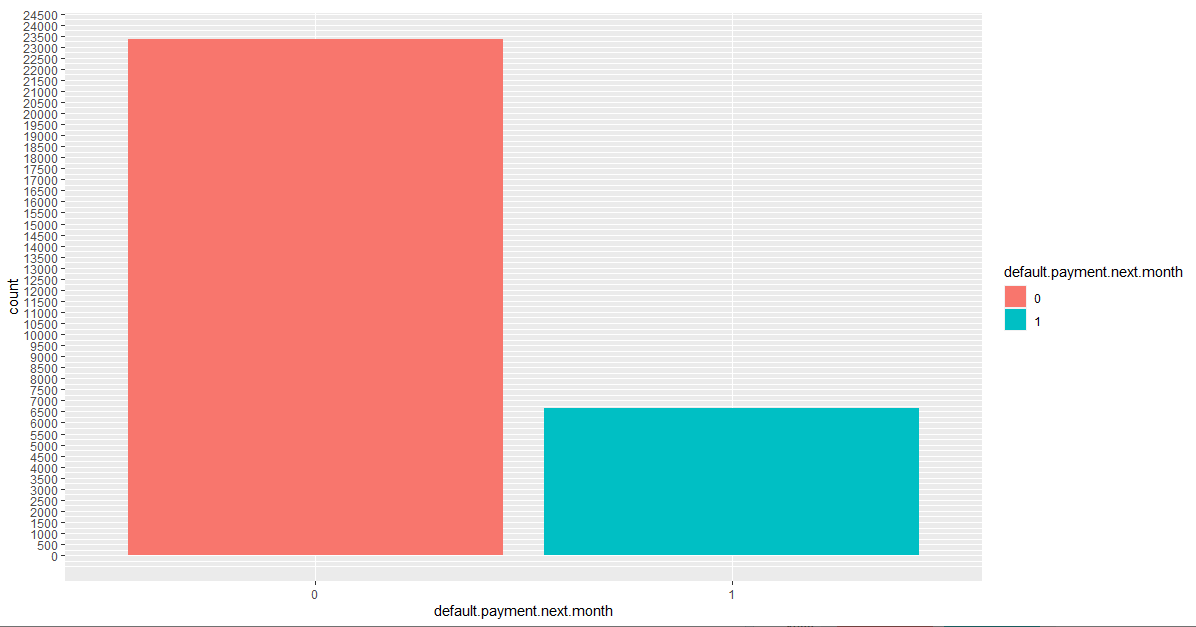


**ניתן לראות כי בנתונים אין NA.**

**הסרת עמודות במודל:**

עמודת "id" הוסרה וזאת מכיוון שעמודה זו לא תורמת לחיזוי משתנה המטרה ואף יכולה לגרום למצב של .overfitting

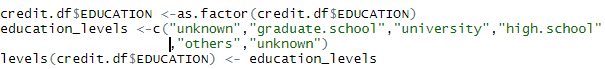
אחרי שראינו כי אין חסרים ברשומות ומחקנו רשומות שלא תורמות למודל, עברנו רשומה רשומה על מנת להבין איך המידע מתנהג, בדוח זה נציג רק את השינויים העיקריים שביצענו.



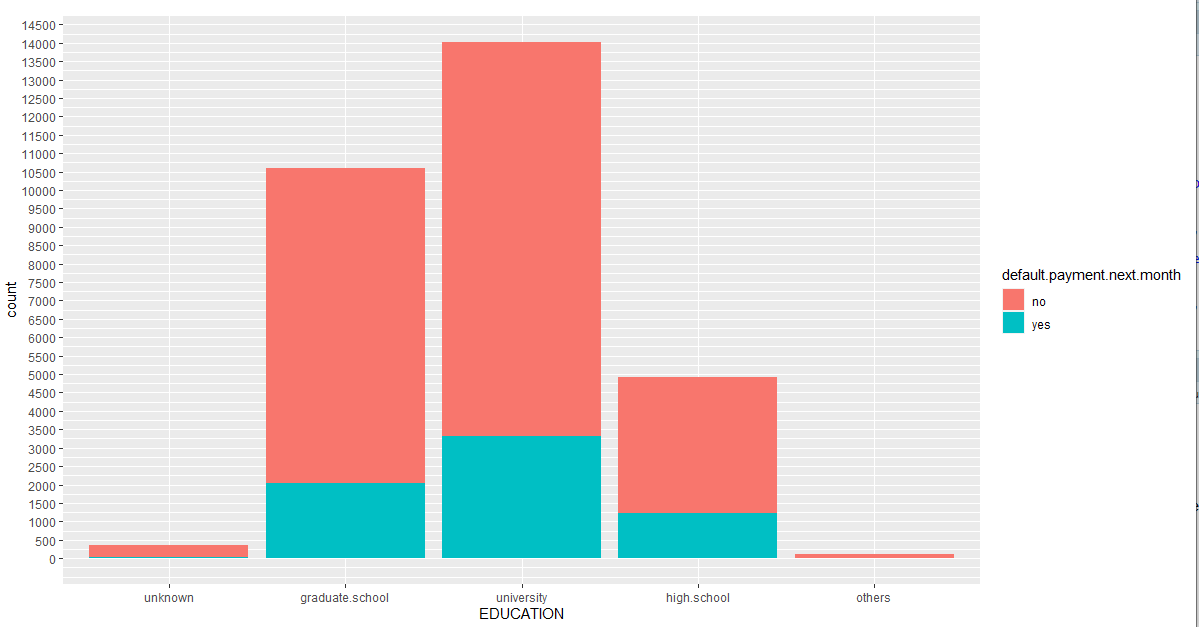
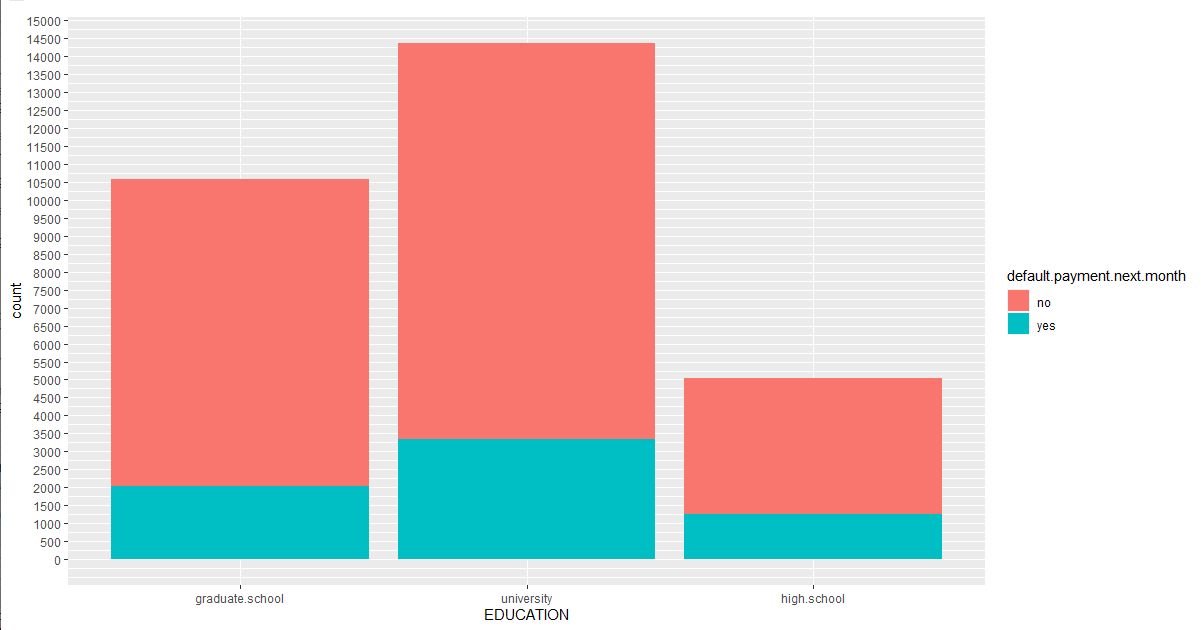
**התחלקות הנתונים** :

ניתן לראות כי מסד הנתונים לא מאוזן ויש יותר תוצאות של אנשים שהחזירו את ההלוואה, לפני הרצת המודלים **נצטרך לטפל בחוסר האיזון** הזה.

כעת עברנו לרשומת הבאה - השכלה הוא משתנה נומרי, הפכנו אותו לפקטור בעזרת הפונקציה הבאה:



לאחר מכן הרצנו היסטוגרם על התוצאות:

ניתן לראות כי מקטגוריות "others" ו"unknown" יש מעט מאוד נתונים (וניתן לראות כי הנתונים לא בהכרח משפיעים על פונקציית המטרה אלא מתפלגים באותה צורה פחות או יותר).

טיפול – נבטל את העמודות האלה ונעמיס אותן על העמודה עם מספר המופעים הכי גדול.

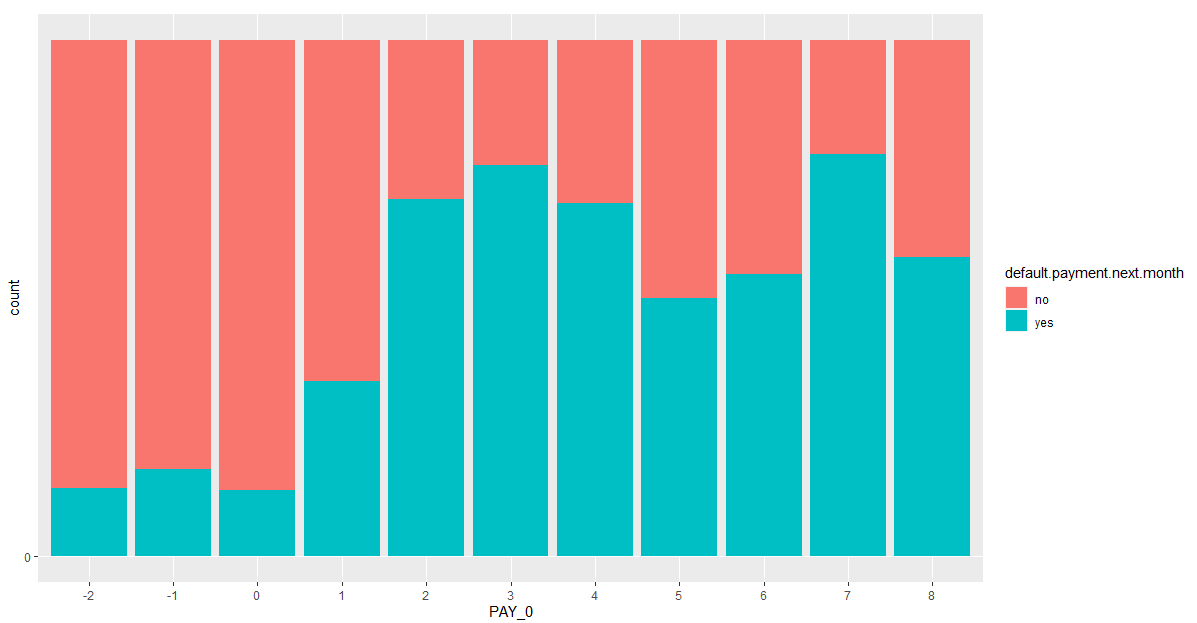
נבחר למחוק את 123 הרשומות עמודת האחר ולהישאר עם משתנה פקטור עם 3 דרגות.

אותו תהליך עשינו עם הלא ידועים ואחרים בקטגוריית הנשואים.

**עבור רשומה כל PAY\_0 עד 6 ביצענו את התהליך הבא:**

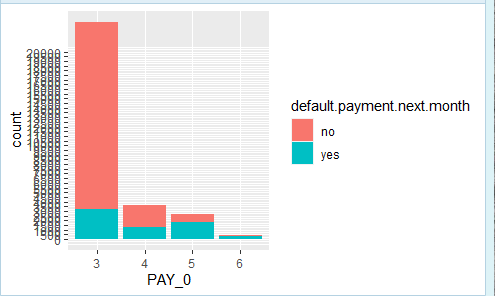
הרצנו היסטוגרם על מנת לראות את ההשפעה של הפקטור על פונקציית המטרה

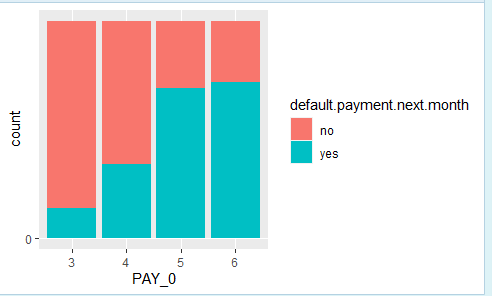
ניתן לראות כי עבור -1 -2 0 אחוזי ההחזרה דומים לעומת זאת עבור תוצאה גדולה מ3 יחסי האי החזרה הם מאוד גבוהים לעומת זאת מספר המופעים שלהם הוא מאוד קטן לכן נחבר את כל האי החזרות מעל 3 ל6 ונחבר את מי שמחזיר (קטן מ-0 ל-3).



3 – החזיר בזמן 5- איחר בחודשיים

4 – איחר בחודש 1 6 – איחר בשלושה חודשים או יותר

תוצאה:

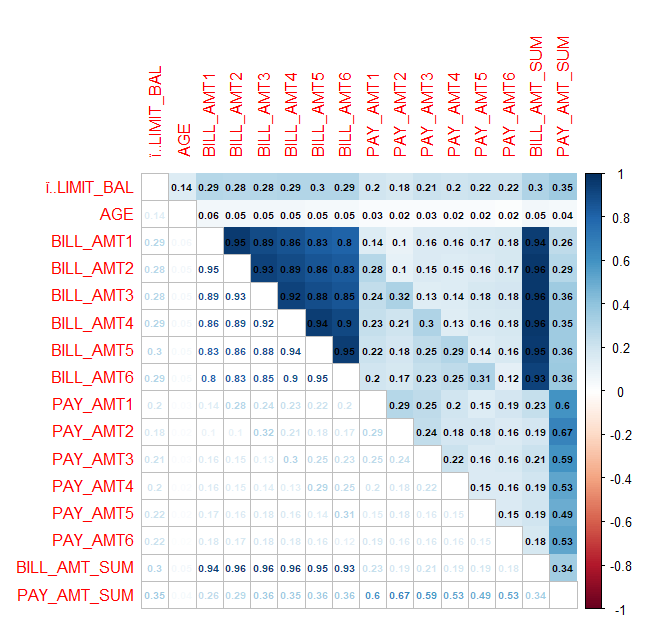


**כעת עברנו לבדיקת העמודות הנומריות:**

2

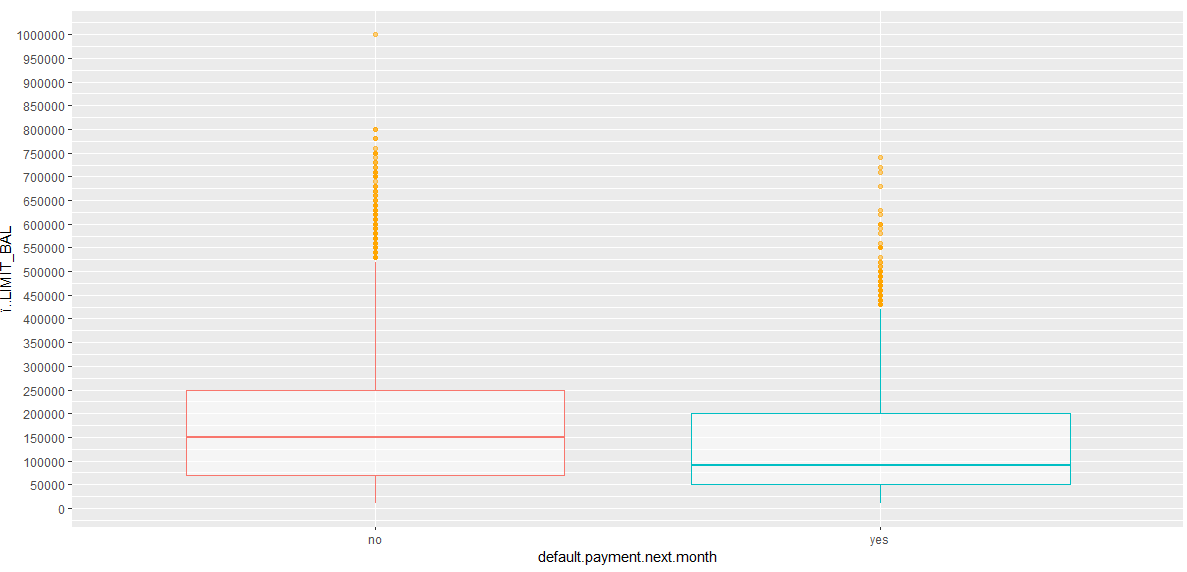
3

1



ניתן לראות כי יש קורלציה דיי חזקה בין גובה החוב לבין גובה החוב בחודשים קודמים (BILL\_AMT1 -6)

כמו כן ניתן גם לראות קורלציה במשתנים המהונדסים שלנו בין סכום החוב בחצי שנה האחרונה לבין החוב בכל חודש.

**כעת בדקנו השפעת הנתונים הנומריים על עמודת המטרה:** 

ניתן לראות כי יש קשר קל בין גובה האשראי לבין החזרי ההלוואות (מי שלא מחזרי בממוצע מסגרת האשראי שלו יותר גבוהה).

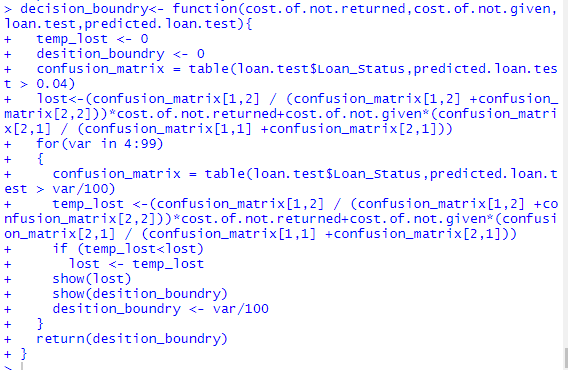
ביצענו ניתוח זה עבור כל פרמטרים של החוב ושל רמת התשלום החודשי, ניתן לראות קשר קל בין אי החזרה לבין גובה החוב.

גובה אשראי לעומת החזר הלוואה

**תהליך הרצת המודלים:**

על מנת לבצע תהליך נכון ואופטימאלי להרצת המודלים עבור כל מודל נבצע:

הרצת המודל על איזון נתונים (oversampaling under samplaing both) נבחר את המוצלח ביותר

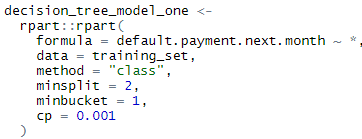
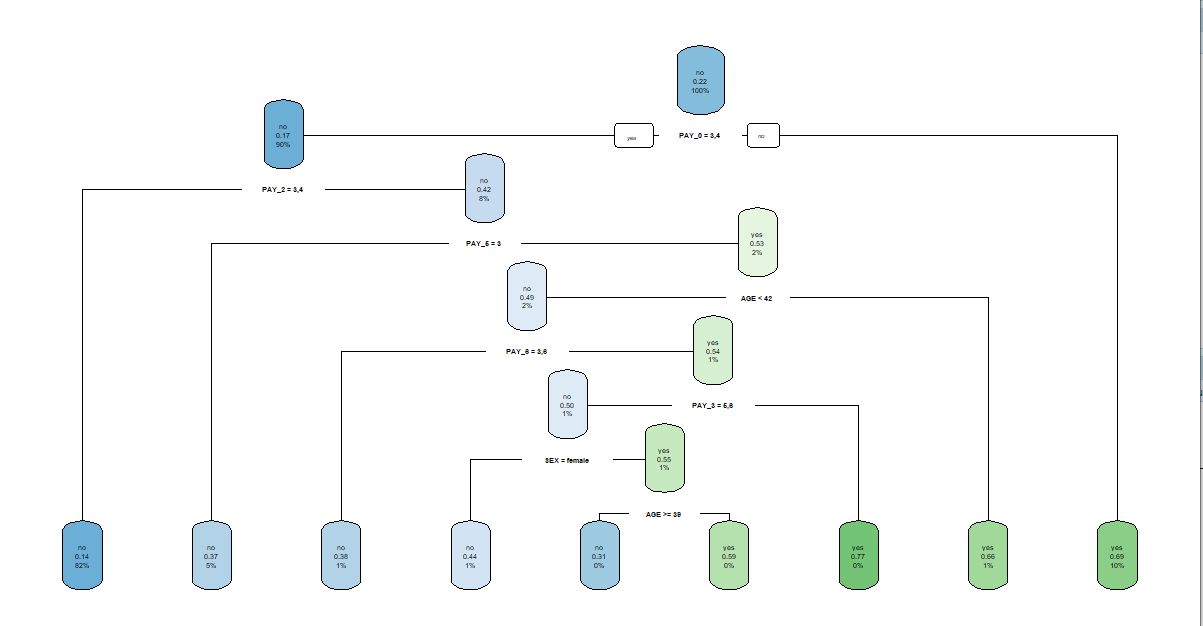
לאחר מכן נריץ על התוצאות את פונקציית הoptimal decision boundary 

על מנת שנקבע את גבול ההחלטה האופטימאלי למודל עבור המדדים הבאים: accuracy, recall, precision

**מטרת הפרויקט: זיהויי האנשים אשר לא יעמדו במסגרת האשראי בחודש הקרוב!**

**מודל עץ החלטות (decision tree):**

המודל הראשון שהרצנו על הנתונים המאוזנים הוא מודל עץ החלטות מודל זה דומה מאוד לתהליך ההחלטה שמבצע בן אדם בשימוש ב"תורת ההחלטות" את המודל ביצענו עם הפרמטרים הבאים:

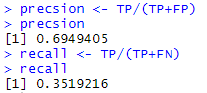
****

הפרמטר המשמעותי ביותר היה CP (complexity parameter) אשר נועד לחסוך זמן איבוד עבור החלטות שלא משנות את התאמת המודל ביותר מערך הCP.

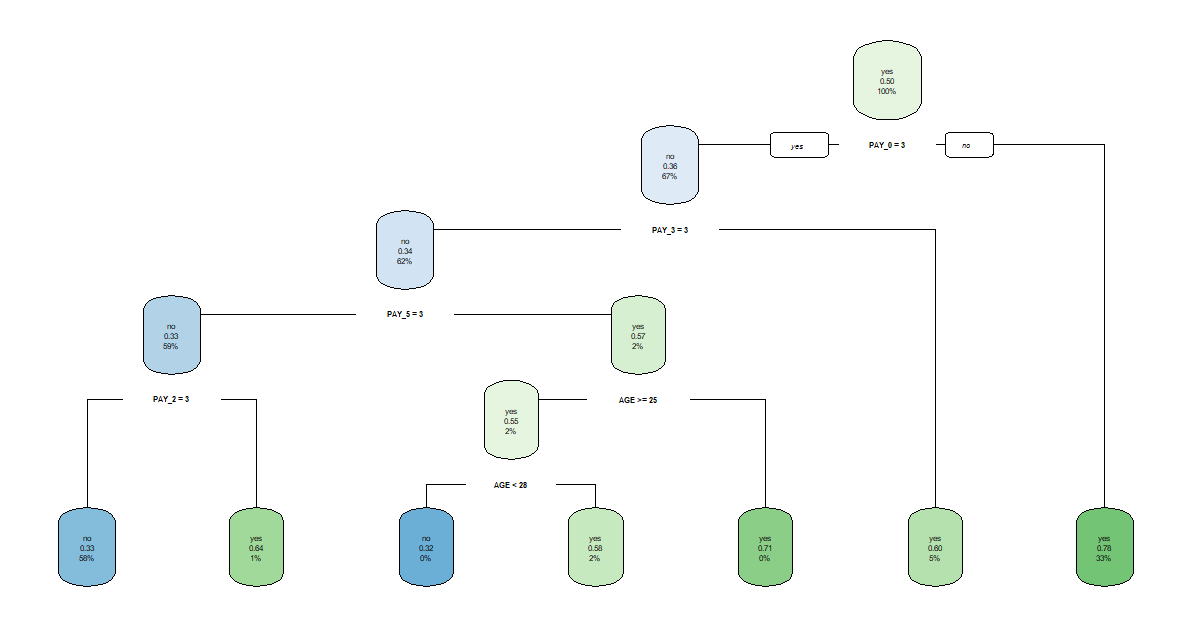
**תוצאות ההרצה של המודל לא מאוזן**:

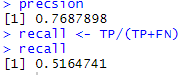
הפרמטר המשמעותי ביותר היה CP (complexity parameter) אשר נועד לחסוך זמן איבוד עבור החלטות שלא משנות את התאמת המודל ביותר מערך הCP.

**ניתן לראות כי תוצאות המודל אינן מיטביות.**

****

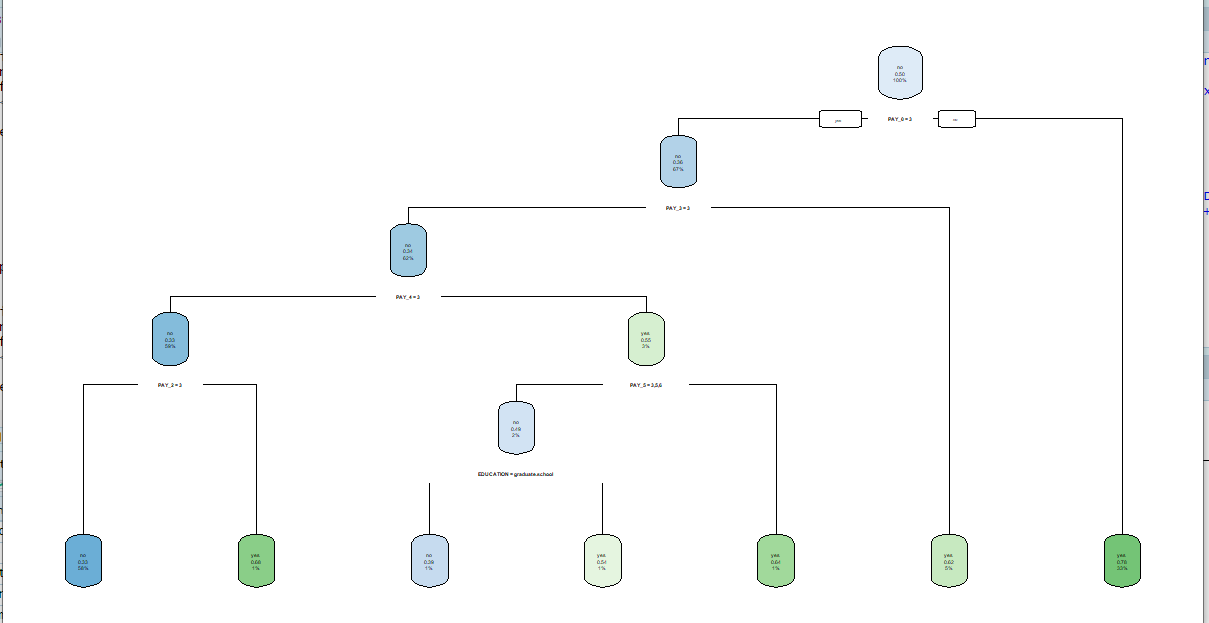
תוצאות ההרצה של המודל oversamp ling:

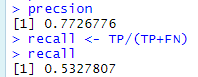


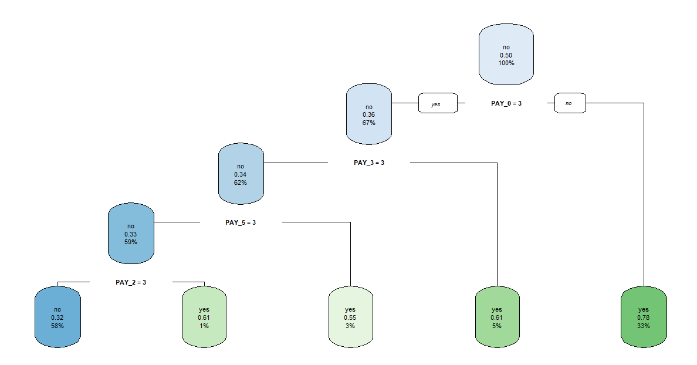


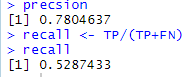
ניתן לראות כי איזון הנתונים אכן שינה את התוצאות עבור עץ החלטות נראה גם את - under sample and both

under sampling:





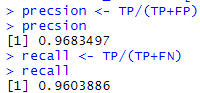
Both:





ניתן לראות כי איזון הנתונים אכן טרם מעט לשיפור הrecall והprecision- של המודל ל- 0.528 ו- 0.78 בהתאמה (בשאר המודלים נראה רק את התוצאה הטובה ביותר) רמת דיוק עומדת עך 0.693.

**מודל בייס נאיבי (naïve base):**





מהנתונים ניתן לראות כי מודל זה אשר בוצע עם התאמת נתונים מסוג (BOTH) מתוך מי שלא החזיר הלוואה המודל הצליח לזהות הצלחה ב96% (!) ומתוך מי שהוא זיהה כי לא יחזיר הלוואה 96.8% אכן לא החזירו, רמת הדיוק הכללית של המודל היא 96%.

**מודל רגרסיה לוגיסטית (logistic regration):**

הרצנו מבחן אנובה לבדיקת הקשר בין המשתנים לעמודת המטרה:

NULL 10622 14727

ï..LIMIT\_BAL 1 434.62 10621 14292 < 2.2e-16

SEX 1 10.65 10620 14281 0.0011007

EDUCATION 2 5.07 10618 14276 0.0790777

MARRIAGE 1 40.19 10617 14236 2.310e-10

AGE 1 0.95 10616 14235 0.3294902

PAY\_0 3 1757.50 10613 12478 < 2.2e-16

PAY\_2 3 87.38 10610 12390 < 2.2e-16

PAY\_3 3 70.48 10607 12320 3.377e-15

PAY\_4 2 34.39 10605 12285 3.400e-08

PAY\_5 3 30.62 10602 12255 1.022e-06

PAY\_6 3 18.05 10599 12237 0.0004298

BILL\_AMT1 1 0.45 10598 12236 0.5018862

BILL\_AMT2 1 1.53 10597 12235 0.2159823

BILL\_AMT3 1 0.24 10596 12234 0.6257701

BILL\_AMT4 1 3.50 10595 12231 0.0614032

BILL\_AMT5 1 1.05 10594 12230 0.3058975

BILL\_AMT6 1 0.03 10593 12230 0.8645876

PAY\_AMT1 1 52.97 10592 12177 3.393e-13

PAY\_AMT2 1 7.61 10591 12169 0.0057982

PAY\_AMT3 1 2.02 10590 12167 0.1547538

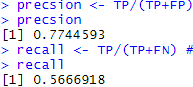
PAY\_AMT4 1 1.44 10589 12166 0.2307632

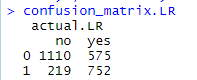
PAY\_AMT5 1 0.32 10588 12166 0.5691160

PAY\_AMT6 1 3.50 10587 12162 0.0612502

BILL\_AMT\_SUM 0 0.00 10587 12162

PAY\_AMT\_SUM 0 0.00 10587 12162 NULL



**

מהנתונים ניתן לראות כי מתוך מי שלא החזיר הלוואה המודל הצליח לזהות הצלחה ב56% ומתוך מי שהוא זיהה כי לא יחזיר הלוואה 77.4% אכן לא החזירו.

ï..LIMIT\_BAL \*\*\*

SEX \*\*

EDUCATION .

MARRIAGE \*\*\*

AGE

PAY\_0 \*\*\*

תוצאות המודל

PAY\_2 \*\*\*

PAY\_3 \*\*\*

PAY\_4 \*\*\*

PAY\_5 \*\*\*

PAY\_6 \*\*\*

BILL\_AMT1

BILL\_AMT2

BILL\_AMT3

BILL\_AMT4 .

BILL\_AMT5

BILL\_AMT6

PAY\_AMT1 \*\*\*

PAY\_AMT2 \*\*

PAY\_AMT3

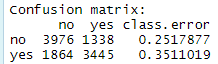
PAY\_AMT4

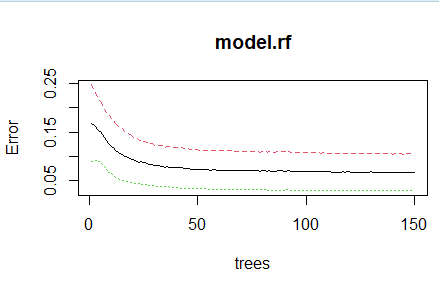
PAY\_AMT5

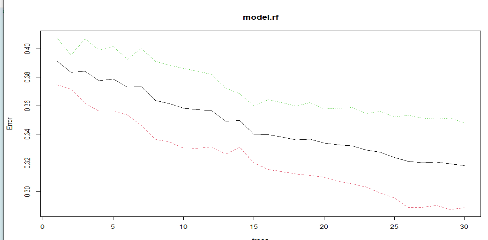
PAY\_AMT6 .

BILL\_AMT\_SUM

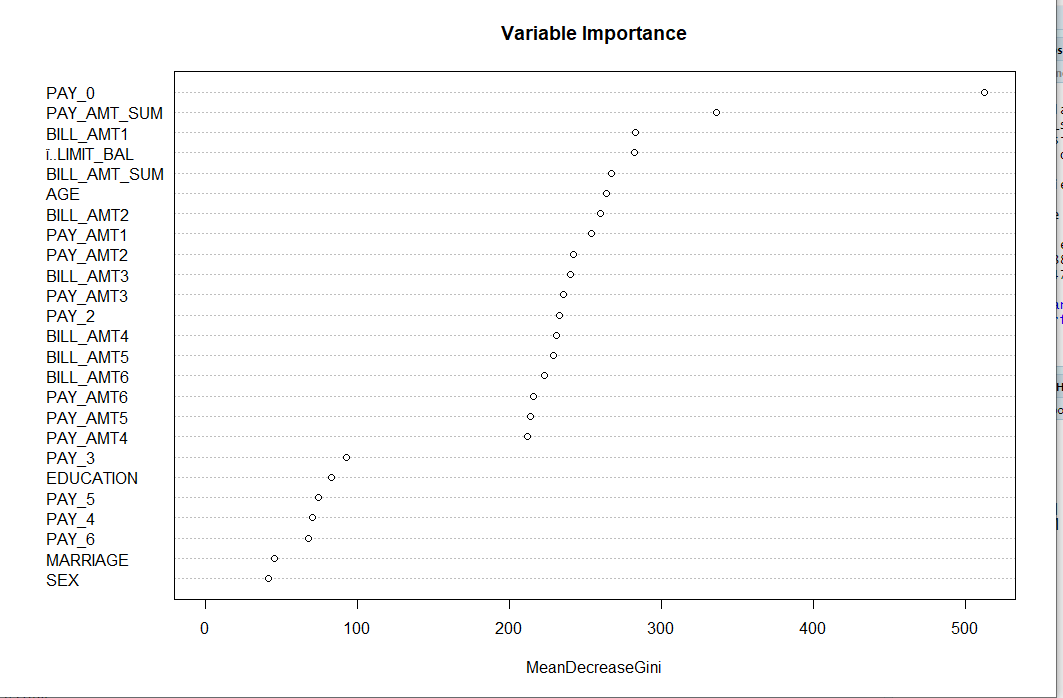
PAY\_AMT\_SUM

**מודל Random forest:**



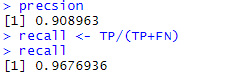


ניתן לראות מהמודל כי אחרי כ 30 עצים רמת הטעות נשארת אותו דבר לכם נבחר להריץ את המודל שוב על 30 עצים.

כעת נבדוק את הפרמטרים בעלי החשיבות הגדולה ביותר:

מהנתונים ניתן לראות כי מתוך מי שלא החזיר הלוואה המודל הצליח לזהות הצלחה ב90% ומתוך מי שהוא זיהה כי לא יחזיר הלוואה 96.7% אכן לא החזירו. רמת הדיוק הכללית ההיא 93.5%

ניתן לראות כי מהנתונים תכונה PAY\_0 (עמידה בהחזרת תשלומים בזמן) היא התכונה המשפיעה ביותר על התוצאה.



תוצאות המודל:

****

**מודל KNN:**

קצת על המודל – מודל KNN הוא ממשפחת המודלים שמשתמשים בsupervised data המודל מבצע קלסיפיקציה על בסיס השכנים של כל מופע מהנתונים.

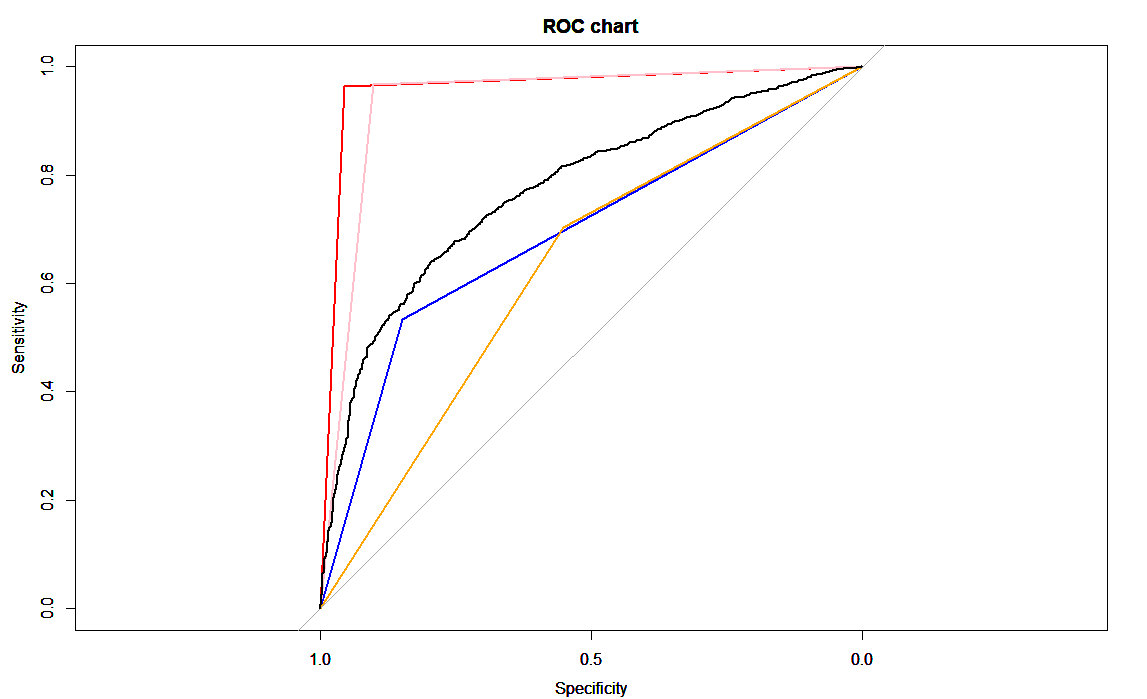
אופן שימוש האלגוריתם :השכנים נלקחים מתוך סדרת אובייקטים של מחלקה (עבור k-NN לסיווג ( או אפיון הערך  
)עבור k-NN לרגרסיה( ידועים.  
חסרון האלגוריתם: חיסרון בולט של האלגוריתם הוא רגישותו למבנה המקומי של הנתונים.

להלן תוצאות המודל**:** במודל שלנו בחרנו לחקור 3 סוגים של ערכים (שכנים ) – כאשר הערך הראשוני הוא שורש מספר המופעים 150, ולאחר מכן הרצנו על K = 100 , 50.

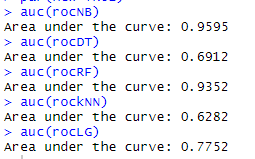
**תוצאות המודל:**

K=150 k=50 k=100

**מסקנה: מבחינת recall עדיף להשתמש בK=150 לעומת זאת מבחינת precision עדיף להשתמש בk=50**



**השוואת מודלים – rock chart:**



**אדום** – Naive Bayes **|**  **כחול**- Decision Tree **|** **שחור** – logistic regression **| ורוד**- RF **|** **כתום** – Knn

**על פי ההשוואה ניתן לראות כי NB הוא המודל המדויק ביותר אשר תופס 95.95 אחוז משטח העקומה. ראינו לפני כן כי הRECALL של המודל הוא 96.8% וה- PRECISION הוא 96.0%**

**וכי מודל random forest תופס כ 93.52% משטח העקומה והצליח במדד הrecall ב96.7% ובprecision**

**מכיוון שאלה המודלים הכי טובים בפער נריץ עליהם את המודל הכלכלי בחלק הבא!**

**מודל כלכלי**

לאחר בדיקה של מודלים והשוואתם אחד לשני הגענו למסקנה שהמודלים הכי מדויקים למקרה שלנו הם naïve base וrandom forest .

מכיוון שמטרת הפרויקט היא להביא ערך כלכלי לחברה על ידי למידת מכונה בנינו מודל כלכלי הממחיש את עדיפות השימוש במודלים לעומת נתינת אשראי שרירותית:

מטרתנו היא לזהות את מי שלא יעמוד במסגרת האשראי! ולכן ניתן לזה עלות גבוהה לעומת מי שכן יחזיר הלוואה:

עלות אי החזר: 1000 $

רווח מהחזר: 300$

(באופן אינסטינקטיבי בעלות אי החזרה גבוהה נעדיף את המודל עם הrecall היותר גבוהה)

תוצאות המודל הכלכלי:

עבור מקרה שרירותי: שבו ניתן הלוואה לכל מי שמבקש נקבל רווח כולל : 373200$ מ30000 מבקשי אשראי שניתן להם



במקרה שנבחר להשתמש באחד המודלים אנחנו ניתן את אשראי רק למי שהמודל חזה שיחזיר את האשראי ונחשב את הרווח עבור כל הTRUE\_POSITIVE וההפסד עבר הFALSE\_POSITIVE או במילים אחרות מדד הRECALL מייצג את אחוז הנכונים והמשלים שלו את ההפסד.

עבור אלגוריתם random forest:



כאשר no מייצג את מי שכן יעמוד בהחזר האשראי וyes את מי שלא יעמוד בהחזרים

מכאן נלווה ל22816 איש כאשר 96.7 אחוז מהאנשים יעמדו בהחזר האשראי ו3.2 אחוז לא יעמדו בהחזרים

22816\*96.7\*300 – הכנסות צפויות

22816\*3.3\*1000 – הפסדים צפויים

סה"כ רווח = 586,599,360$

תוצאות אלא מראות על פער עצום ברווחים למרות שניתנן פחות אשראי מה שמראה את חשיבותו של מודל חיזוי טוב הפער הוא מעל חצי מיליארד דולר ברווחים!!!

עבור אלגוריתם naïve base:



כאשר no מייצג את מי שכן יעמוד בהחזר האשראי וyes את מי שלא יעמוד בהחזר

מכאן נעמיד אשראי ל22816 איש כאשר 96.8 אחוז מהם יעמדו בהחזר ו3.2 אחוז לא יעמדו

22510\*96.8\*300 – הכנסות צפויות

22510\*3.2\*1000 – הפסדים צפויים

סה"כ רווח = $646,656,025

**המלצות:**

המלצתנו לחברת האשראי להשתמש בחיזוי של מודל בייס נאיבי, מודל זה ישפר את הרווחים שלהם על פני מתן אשראי שרירותי במעל 600 מיליון דולר!! יאפשר להם לתת אשראי זול יותר לזכאים ולצמצם את הוצאות האכיפה שלהם.

**מסקנות:**

אחרי ביצוע הפרויקט ראינו כי באמצעות כלים חינמיים ויחסית פשוטים ניתן לתת ערך כלכלי עצום לחברות אשר יש להן דאטה מתאים על הלקוחות שלהן.

ביצענו ניתוח על פי 5 מודלים ועבור כל מודל ביצענו העצמה לנתונים כך שיאזן את כמות המחזירים מול הלא מחזירים התוצאות והתגובה ל המודלים הן מדהימות ומפתיעות.

הצלחנו להגיע לתוצאות דיוק של מעל 95% בשתי מודלים שונים.

את התועלת מביצוע למידת מכונה על דאטה מתאימה ניתן לראות על פי המודל הכלכלי שהצגנו שמראה פערים עצומים ברווחי החברות (החזויים) בין התנהלות עם מודל לחיזוי, לבין התנהלות ללא מודל לחיזוי, מה שממחיש את האמירה שכל ארגון חייב אסטרטגיה לשילוב למידת מכונה בשנת 2020.

מסקנתנו מעבודה זו היא שארגון שלא יעשה שימוש על הדאטה שלו בעזרת למידת מכונה עלול להישאר מאחור ולהיות לא רלוונטי מול המתחרים אשר ישתמשו במודלי חיזוי.