1. הגדרת הבעיה והכנת הנתונים

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **שם תכונה** | **תיאור התכונה** | **סוג** | **יחידות מדידה** | **תחום ערכים** | **ממוצע** | **סטיית תקן** | **ערכים חסרים** | **ערכים לא חוקיים** |
| age | גיל | נומרי - שלם | שנים | (17)-(98) | 40.02 | 10.42 | אין | אין |
| job | סוג המשרה | קטגורי |  | 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', student', 'technician', 'unemployed', 'unknown' |  |  | מסומנים בUnknow כמות : 330 | אין |
| marital | מצב משפחתי | קטגורי |  | 'divorced' - גרוש או אלמן 'married', 'single', 'unknown' |  |  | מסומנים בUnknow כמות : 80 | אין |
| education | השכלה | קטגורי |  | basic.4y', 'basic.6y', 'basic.9y', 'high.school',  'illiterate', 'professional.course', 'university.degree', 'unknown' |  |  | מסומנים בUnknow כמות : 1731 | אין |
| default | פושט רגל | קטגורי |  | 'no','yes','unknown' |  |  | מסומנים בUnknow כמות : 8597 | אין |
| housing | משכנתא | קטגורי |  | 'no','yes','unknown' |  |  | מסומנים בUnknow כמות : 990 | אין |
| loan | הלוואה אישית | קטגורי |  | 'no','yes','unknown' |  |  | מסומנים בUnknow כמות : 990 | אין |
| contact | אמצעי יצירת קשר | קטגורי |  | 'cellular','telephone' |  |  | אין | אין |
| month | חודש אחרון בו נוצר קשר | קטגורי |  | 'jan', 'feb', 'mar', ..., 'nov', 'dec' |  |  | אין | אין |
| day\_of\_week | יום בשבוע אחרון בו נוצר קשר | קטגורי |  | 'mon','tue','wed','thu','fri' |  |  | אין | אין |
| duration | משך שיחה אחרונה | נומרי - שלם | שניות | (0)-(4918) | 258.29 | 259.28 | אין | אין |
| campaign | כמות התקשרויות בקמפיין הנוכחי | נומרי - שלם | יח' | (1)-(56) | 2.57 | 2.77 | אין | אין |
| pdays | כמות ימים מאז התקשרות בקמפיין קודם | נומרי - שלם | ימים | (0)-(27), 999 - לא נוצר בעבר קשר | 6.01 | 3.82 | אין | אין |
| previous | כמות התקשרויות לפני קמפיין זה | נומרי - שלם | יח' | (0)-(7) | 0.17 | 0.49 | אין | אין |
| poutcome | תוצאת הקמפיין הקודם | קטגורי |  | 'failure','nonexistent','success' |  |  | אין | אין |
| emp.var.rate | שינוי המועסקים במשק (רבעוני) | נומרי - רציף | אחוזים | (-3.4) - (1.4) | 0.08 | 1.57 | אין | אין |
| cons.price.idx | מדד המחירים לצרכן (חודשי) | נומרי - רציף | יח' | (92.201) - (94.767) | 93.58 | 0.58 | אין | אין |
| cons.conf.idx | מדד הבטחון הכלכלי (חודשי) | נומרי - רציף | יח' | (-50.8) - (-26.9) | -40.50 | 4.63 | אין | אין |
| euribor3m | ריבית היוריבור ל3 חודשים (יומי) | נומרי - רציף | אחוזים | (0.634) - (5.045) | 3.62 | 1.73 | אין | אין |
| nr.employed | כמות המועסקים (רבעוני) | נומרי - רציף | יח' | (4963.6) - (5228.1) | 5167.04 | 72.25 | אין | אין |
| y | משתנה יעד, האם חתם על פקדון | בינארי |  | 'yes', 'no' |  |  | אין | אין |

סה"כ רשומות לפני הניקוי 41189

סה"כ רשומות לאחר הניקוי הנתונים (ראו סעיף ה'): 40119

1. **מטרת כריית המידע**: לחזות האם לקוח יחתום על תנאי הפקדון שהוצעו לו במהלך שיחות השיווק הטלפוני
2. **שלבי ה-KDD :**
3. איסוף ושמירת הנתונים

בשלב זה, נבחר מה הוא סט הנתונים עליו נעבוד, לרוב, משיקולי יעילות וביצועים נעדיף סטים קטנים של מידע ומשיקולי איכות נעדיף סטים גדולים של מידע.

בנוסף, לרוב קיימים שיקולים של בחירת תקופות במיוחד בלמידה מסוג time series prediction. במקרה זה, נשקול על איזו תקופה נרצה ללמוד ככל שתהיה ישנה יותר תהיה פחות רלוונטית (לרוב) אך יותר מלאה.

במקרה זה, נבחר את הקובץ המלא ביותר והמקיף ביותר (לפי כמות המאפיינים) - bank-additional-full

בנוסף, לעיתים בשלב זה כאשר ערך הTarget או Class מסויים אינו מיוצג בהתאם למציאות או שהינו נדיר בסט הנתונים שלנו, נוכל להשתמש בשיטות של Under/over sampling של המידע כדי לא יצור הטיות באלגוריתמים מסויימים

1. ניקוי הנתונים (טיפול בערכים חסרים)

נחפש רשומות בעלות ערכים לא הגיוניים, לדוגמא ערכים מספריים חריגים לפי מינימום ומקסימום של השדה עבור שורות חריגות נחליט אם להשמיט אותן לגמרי (במידה ומדובר בכמות קטנה) , לתקן את הערכים או לשים שם ערך ריק.

עבור ערכים ריקים אנו יכולים לבחור לטפל בהם לפי מס' שיטות : השלמה ידנית, קבוע כלשהו, הממוצע, ממוצע של התצפיות ששייכות לאותה קטגוריה, השלמה ע"י למידה מונחית.

נשים לב כי ישנם אלגוריתמים אשר יודעים להתמודד עם ערכים חסרים (עץ החלטה) וישנם כאלה שלא (

1. ביצוע טרנספורמציות על הנתונים

בשלב זה נבחן האם ניתן לבצע שינוים על המידע אשר יתרמו למודל בין אם בשינויי המאפיינים הקיימים או ביצירת מאפיינים חדשים, דוגמא אחת היא הפיכת משתנה כמותי לאינדיקטור (לדוגמא, הפיכת משתנה של כמות רכישות לאינדיקטור של האם ביצע רכישה), ביצוע אגרגציות, ביצוע נורמליזציה של משתנה קיים ועוד.

1. ביצוע כריית המידע

בשלב זה נחליט על סוג כריית המידע הנדרש לפרוייקט הספציפי הזה בין האפשרויות הקיימות של classification, prediction, clustering ועוד.

נחליט על אופן חלוקת המידע לtrain \ test כאשר על הtrain data ייבנה המודל וייבדק על הtest data לפי קריטריון שנבחר.

בנוסף, נבחר אלגוריתם אחד או כמה מתוך המשפחה שבחרנו לעיל ונבנה מודל בעזרתו.

לרוב לאחר הרצה האלגוריתמים נגלה כי אנו נדרשים לבצע התאמות ושינויים הן במאפיינים שבחרנו והן בפרמטרים של האלגוריתמים עצמם.

ייתכן ונצטרך להתמודד עם בעיות שונות כמו overfitting ולשנות את האלגוריתם שלנו (לדוגמא pre\post pruning בעץ)

לבסוף, נבחר את המודל שנותן את התוצאות הטובות ביותר עבורנו ולפי הקריטריון שהגדרנו.

1. ידע

את המודל שקבלנו נפעיל על מידע חדש (unseen data) אשר אין אנו יודעים את תוצאת הclassification / prediction / clustering שלו ונקבל חיזוי מהמודל. לפי סטטיסטיקות שונות הנובעות בין היתר מהconfusion metrix שנקבל מהמודל נוכל להעריך את מידע הדיוק של החיזוי שלנו.

### חלופות אפשריות לביצוע כריית מידע

### עץ החלטה ID3

ID3 הוא אלגוריתם לבניית עץ החלטה, אלגוריתם זה, מתחיל בראש העץ עם הסט המלא של הנתונים, עבור סט זה מצבע איטרציות על כל המאפיינים ומחשב את האנטרופיה או הinformation gain על הסט בהתאם למאפיין. עבור המאפיין בעל האנטרופיה הנמוכה ביותר או הinformation gain הגבוה ביותר הענף מפוצל בנק' מסויימת. וכך נוצרים ענפים חדשים. ברגע שהעץ לא יכול להתפצל יותר (לדוגמא כאשר כל הערכים שייכים לclass מסוים \ לא נשארו עוד מאפיינים ועוד') נקבע כי תת הסט הנוכחי הוא עלה.

אלגוריתם זה אינו מבצע pruning ולכן במקרים רבים נגרם מצב של overfitting.

בנוסף, אלגוריתם זה אינו מבטיח פתרון אופטימלי שכן הינו אלגוריתם חמדני.

### עץ החלטה C4.5

C4.5 הינו שיפור של ID3 אשר עובד בצורה דומה (פיצולים של העץ לפי IG \ Entropy). השיפורים הם :

1. התמודדות עם ערכים רציפים ובדידים כאחד
2. התמודדות עם ערכים חסרים
3. ביצוע pruning לעץ

במקרה שלנו ישנם גם ערכים רציפים וגם בדידים וכן ערכים חסרים, בנוסף ביצוע pruning עוזר להתמודדות עם overfitting ולכן נעדיף אלגוריתם זה על פני ID3

### רגרסיה לינארית מרובת משתנים

שיטה שבה ממודל הקשר בין משתנה תלוי לבין משתנים מסבירים. לפי מודל זה, נבנית נוסחא לניבוי של משתנה המטרה. עבור כל משתנה שאינו בנוסחא, נבחר במשתנה אשר תורם לירידה בסכום הריבועים של השגיאה בצורה המקסימלית.

השיטה מניחה כי קיים קשר לינארי בין המשתנה התלוי למשתנים המסבירים, המשתנים מתפלגים נורמאלית, הומוסקדסטיות בין המשתנים.

במקרה שלנו, אין אנו יודעים על קיומו של קשר לינארי (לדוגמא ייתכן שחותמים על תנאי פיקדון יותר בשנות ה20 וה60 ופחות בשנות ה30-50 בחיים ולכן אי אפשר לייצר מקדם שמיש) וכן המשתנים אינם מתפלגים נורמאלית ולכן נעדיף שלא להשתמש בשיטה זו.

### עץ החלטה CART מבוסס מדד GINI

CART היא **קבוצה** של עצי החלטה. נשתמש בעץ החלטה המשתמש במדד ג'יני בקביעת הפיצול.

הבדלים עיקריים בין מדד GINI ל Entropy :

1. ג'יני מיועד למאפיינים רציפים ואנטרופיה למאפיינים קטגוריאליים ([מקור](http://paginas.fe.up.pt/~ec/files_1011/week%2008%20-%20Decision%20Trees.pdf))
2. ג'יני יעדיף למצוא קטגוריות גדולות ביותר ואנטרופיה נוטה לקטגוריות קרובות ל50% מהמידע ([מקור](http://paginas.fe.up.pt/~ec/files_1011/week%2008%20-%20Decision%20Trees.pdf))
3. המדדים נותנים תוצאות שונות בפחות מ2% מהמקרים ([מקור](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.57.9764&rep=rep1&type=pdf))

ייתכן ובמקרה שלנו מדד ג'יני יהיה עדיף על אנטרופיה וייתכן שלהיפך. נבחן את שניהם.

# שלבי הכנת הנתונים

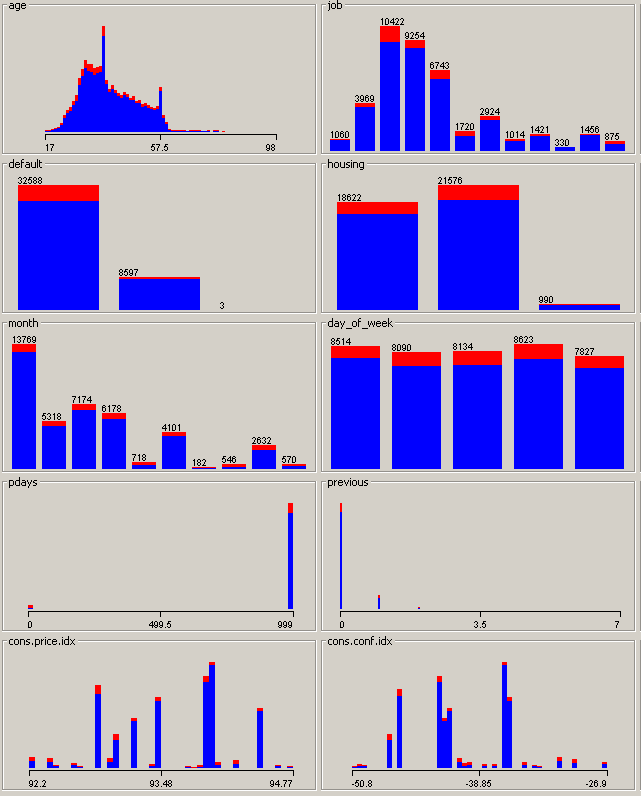
**כפי שצוין לעיל, נבחר בקובץ המלא והמקיף ביותר :** bank-additional-full

במידע שלפנינו לא נמצאו ערכים בלתי חוקיים.

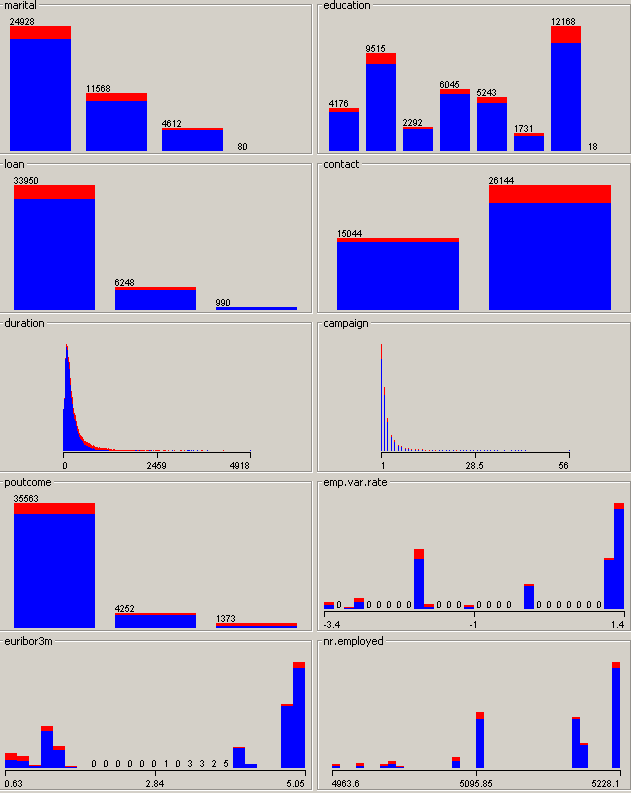
נמצאו מספר עמודות עם ערכים חסרים כפי שצוין בסעיף א'

בערכים חסרים ניתן לטפל במספר שיטות כפי שצוין בסעיף ג'

אופן הטיפול עבור כל מאפיין יוסבר בהמשך.



נציג את הנתונים בצורה גראפית :



|  |  |
| --- | --- |
| **תחום ערכים** | **שם תכונה** |
| לא נמצאו ערכים חריגים או חסרים, נשאיר את כל השורות | age |
| לא נמצאו ערכים חריגים, נמצאו 330 ערכים חסרים אשר מהווים 0.8% מהמידע, התפלגות הערכים נראית סבירה, לאור הכמות הנמוכה, ניתן להשמיט שורות עם ערכים חסרים אלו | job |
| לא נמצאו ערכים חריגים, נמצאו 80 ערכים חסרים אשר מהווים 0.2% מהמידע, התפלגות הערכים נראית סבירה, לאור הכמות הנמוכה, ניתן להשמיט שורות עם ערכים חסרים אלו | marital |
| לא נמצאו ערכים חריגים, נמצאו 1731 ערכים חסרים אשר מהווים 4.2% מהמידע ובתוכם הן תוצאות "yes" ו"no", התפלגות הערכים נראית סבירה.  כמות הערכים החסרים אינה מזערית וכן השלמה לפי רוב הינה בעייתי שכן ייצר BIAS לטובת הקבוצה הגדולה ביותר וכן כלל לא סביר כי רוב הערכים הריקים הם של קבוצה אחת אזי נחלק לפי תוחלת בין כל הקבוצות. | education |
| קיים רק 3 "yes" כל השאר ריקים או "no", נמצאו 8597 ערכים חסרים אשר מהווים כ20% מהמידע ובתוכם הן תוצאות "כן" ו"לא"  לאור 2 הסיבות לעיל נעדיף לא להשתמש במאפיין זה כלל שכן נראה כי אינו אינפורמטיבי כלל. | default |
| לא נמצאו ערכים חריגים, נמצאו 990 ערכים חסרים אשר מהווים 2.4% מהמידע, התפלגות הערכים נראית סבירה, לאור הכמות הנמוכה, ניתן להשמיט שורות עם ערכים חסרים אלו | housing |
| לא נמצאו ערכים חריגים, נמצאו 990 ערכים חסרים אשר מהווים 2.4% מהמידע (חופפים עם שורות הhouseing הריקות), התפלגות הערכים נראית סבירה  לאור הכמות הנמוכה, ניתן להשמיט שורות עם ערכים חסרים אלו | loan |
| לא נמצאו ערכים חריגים או חסרים, נשאיר את כל השורות | contact |
| בחודש מאי מבוצעות משמעותית יותר הצעות מחודשים אחרים, כנראה שיש סיבה עסקית לכך אולם ויזואלית נראה כי שונות אחוז ה”yes” החודשי הוא סביר.  פרט לכך, אין ערכים חסרים או חריגים, נשאיר את כל השורות. | month |
| לא נמצאו ערכים חריגים או חסרים, נשאיר את כל השורות | day\_of\_week |
| בקובץ ההסבר נאמר כי :  this attribute highly affects the output target )e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known.  נשים לב כי מאפיין זה הינו בעייתי שכן בעולם האמיתי אינו קיים לפני ביצוע שיחה ולכן לא ניתן להשתמש בו לחיזוי.  לאור זאת, נחליט לא להשתמש במאפיין זה כלל לטובת החיזוי | duration |
| ישנה כמות מזערית של תצפיות עם ערכים גדולים מ7 אולם אין צרוך בטיפול מיוחד בהן, סביר כי אין להם ערך יאינפורמטיבי והמודל פשוט לא ישתמש בהן | campaign |
| ל93.8% מהרושומות הערך הינו 999 – קרי, לא נוצר קשר בעבר עם הלקוח.  על מנת שהמסווג ידע לסווג את המידע הזה הערכים צריכים להיות מובחנים מהשאר. לאור זאת שמדובר במידע רציף המסווג יוכל לחלק את טווח הערכים ב >27 לדוגמא (הערך המקסימלי התקין) ובכך להפריד בין לקוחות שפנו אליהם בעבר לכאלו שלא.  נראה כי ללקוחות שפנו אליהם בעבר אחוז ה”yes” הינו גבוה יותר אולם המסווג יחליט האם אינפורמציה זאת תורמת או לא.  נשאיר שדה זה כפי שהוא. | pdays |
| לא נמצאו ערכים חריגים או חסרים, נשאיר את כל השורות | previous |
| לא נמצאו ערכים חריגים או חסרים, נשאיר את כל השורות | poutcome |
| לא נמצאו ערכים חריגים או חסרים, נשאיר את כל השורות | emp.var.rate |
| לא נמצאו ערכים חריגים או חסרים, נשאיר את כל השורות | cons.price.idx |
| לא נמצאו ערכים חריגים או חסרים, נשאיר את כל השורות | cons.conf.idx |
| לא נמצאו ערכים חריגים או חסרים, נשאיר את כל השורות | euribor3m |
| לא נמצאו ערכים חריגים או חסרים, נשאיר את כל השורות | nr.employed |

לאחר ניקוי הנתונים נשארו 40119 רשומות.

# 2) סיווג וחיזוי

1. נבחר בשיטת CART וC4.5

השיטות הוסברו [לעיל](#_חלופות_אפשריות_לביצוע)

לא נבחר ברגרסיה לינארית מרובת משתנים מהסיבות שהוסברו [לעיל](#_רגרסיה_לינארית_מרובת) (בקצרה, לא קיים קשר לינארי, המשתנים לא מתפלגים נורמאלית)

לא נבחר בID3 מהסיבות שהוסברו [לעיל](#_עץ_החלטה_ID3) (בקצרה, C4.5 עדיף עליו בכל המאפיינים)

1. שלבי C4.5 :

ראשית נגדיר את תנאי העצירה :

* כל הערכים בתת העץ שייכים לאותה קטגוריה
* אין מאפיין אשר נותן information gain

שלבי האלגוריתם :

1. התחל מהDATA SET המלא
2. אם הגענו לתנאי עצירה, עצור
3. לכל מאפיין a
   1. מצא את הinformation gain המנורמל בחלוקת הצומת הנוכחי ב a
4. קבע את a\_best כמאפיין עם הinformation gain הגבוה ביותר
5. פצל את הצומת הנוכחי לפי a
6. הרץ שלבים 2-6 על הבנים של הצומת הנוכחי

CART:

שלבי CART דומים מאוד לC4.5 עם כמה הבדלים מהותיים :

1. הפיצולים הנעשים בCART הינם בינאריים לעומת C4.5 היודע לפצל ליותר מקבוצה אחת.
2. CART משתמש במדד ג'יני במקום information gain בסעיפים 3,4 לעיל.
3. CART מבצע גיזום ע"י מודל סיבוכיות-עלות עם פרמטרים הנגזרים מהcross validation לעומת C4.5 המבצע מעבר יחיד על העץ וגוזם לפי איבוד אינפורמציה פוטנציאלי על פי הconfidence level המוגדר.

([מקור](http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf))

1. ביצעתי את הריצות על הקובץ הנקי. בכל שיטה הרצתי מס' פעמים את האלגוריתם כדי לקבל תחושה של השפעת הפרמטרים על התוצאות.

בשתי השיטות השתמשתי ב10-Fold cross validation לבדיקת המודל לפי גישה זו מחלקים את כל סט הנתונים ל-10 קבוצות ומריצים את אחד האלגוריתמים 10 פעמים, כאשר בכל פעם קבוצה אחרת מהווה TestSet ושאר הקבוצות – TrainingSet.

להרצת C4.5 השתמשתי באלג' J48 בWeka ולהרצת CART השתמשתי בSimpleCart

דרשתי בכל מקרה כמות יחסית גדולה של ערכים בעלים, ברירת מחדל של 2 נראית לי נמוכה בצורה קיצונית אשר תגרום לזמני ריצה ארוכים, עצים גדולים ו Overfitting.

להלן התוצאות :

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Run #** | **leaf min** | **Confidence Factor** | **Pruning** | **TP** | **FP** | **TN** | **FN** | **Precision** | **Recall** | **Accuracy** | **Specificity** | **ROC Area** | **Number of Leaves** |
| C4.5 | 1 | 50 | 0.1 | Y | 1013 | 524 | 35074 | 3508 | 65.9% | 22.4% | 89.9% | 98.5% | 70.4% | 24 |
| 2 | 200 | 0.1 | Y | 947 | 458 | 35140 | 3574 | 67.4% | 20.9% | 89.9% | 98.7% | 70.5% | 9 |
| 3 | 50 | 0.25 | Y | 1094 | 555 | 35043 | 3427 | 66.3% | 24.2% | 90.1% | 98.4% | 70.9% | 34 |
| 4 | 200 | 0.25 | Y | 1002 | 477 | 35121 | 3519 | 67.7% | 22.2% | 90.0% | 98.7% | 70.9% | 9 |
| 5 | 50 | 0.5 | Y | 1177 | 627 | 34926 | 3389 | 65.2% | 25.8% | 90.0% | 98.2% | 76.2% | 67 |
| 6 | 200 | 0.5 | Y | 1038 | 485 | 35113 | 3483 | 68.2% | 23.0% | 90.1% | 98.6% | 76.0% | 28 |
| 7 | 50 | - | N | 1175 | 710 | 34888 | 3346 | 62.3% | 26.0% | 89.9% | 98.0% | 79.0% | 141 |
| 8 | 200 | - | N | 1038 | 486 | 35112 | 3483 | 68.1% | 23.0% | 90.1% | 98.6% | 78.1% | 30 |
| CART | 1 | 50 | - | Y | 1014 | 527 | 35071 | 3507 | 65.8% | 22.4% | 89.9% | 98.5% | 70.0% | 19 |
| 2 | 200 | - | Y | 1023 | 501 | 35097 | 3498 | 67.1% | 22.6% | 90.0% | 98.6% | 70.0% | 7 |
| 3 | 50 | - | N | 1160 | 704 | 34894 | 3361 | 62.2% | 25.7% | 89.9% | 98.0% | 70.0% | 166 |
| 4 | 200 | - | N | 1070 | 566 | 35032 | 3451 | 65.4% | 23.7% | 90.0% | 98.4% | 75.8% | 18 |

ד+ה) על מנת להעריך את דיוק השיטה הצגתי בסעיף הקודם שלל פרמטרים להערכת טיב המסווג.

הפרמטרים הם :

* Precision : מידע ה"דייקנות" – כמה תצפיות הן חיוביות מתוך כלל התצפיות שסווגו כחיוביות
* Recall \ Sensitivity: מידע ה"שלמות" – כמה תצפיות חיוביות אכן סווגו ככאלו
* Accuracy : אחוז התצפיות שסווגו נכונה ע"י המסווג
* Specificity: אחוז התצפיות השליליות שסווגו כשליליות.
* ROC Area : השטח תחת עקומת הROC אשר מייצגת את שקלול התמורות שבין שיעור ה True positive לבין שיעור הFalse positive

כעת עלינו להחליט מה הוא המדד אשר משקף נאמנה את איכות המסווג במקרה הספציפי שלנו.

אם כן, מטרת המסווג היא לסווג לקוחות קבוצה אשר תכיל כמה שיותר לקוחות אשר יחתמו על תנאי הפיקדון בשיחה טלפונית.

ככל שיותר אנשים בקבוצה במסווגת יחתמו על תנאי הפיקדון כך איכות המודל תגדל.

אם כך, לדעתי הפרמטר החשוב עבורנו הינו Precision שהרי ככל שמדד זה יותר גבוה, כך אחוז התצפיות השליליות (לקוחות שביצעו להם שיחות ובסוף הם לא חתמו) תרד וכך עבודת המכירות תהיה יעילה יותר.

נוכל כעת להעריך את מידת היעילות של שני השיטות :

1. CART – בשיטה זו נראה כי הדיוק המרבי הושג ע"י ריצה מס' 2: 67.1% דיוק. בריצה זו השתמשנו בהגדרה של מינימום 200 תצפיות בעלה.
2. C4.5 – בשיטה זו נראה כי הדיוק המרבי הושג ע"י ריצה 6 בהפרש של 0.1 מריצה 8, 2 הריצות מוגדרות למינימום 200 תצפיות בעלה וללא גיזום או גיזום מזערי בלבד.

מסקנות כלליות מהריצות :

1. ריצה עם דרישה לכמות גדולה יותר של ערכים בעלים כאשר כל השאר קבוע, מובילה לתוצאות טובות יותר במדד הדיוק שבחרנו וזאת בגלל שנוצר עץ קטן יותר עם סיכוי נמוך יותר לoverfitting. (הדרישה היא של 50 לעומת 200, סביר כי בכמות גדולה מדי יצא עץ קטן מדי אשר לא יסווג היטב, בדקתי אמפירית ונראה כי סביבות 250 תצפיות בעלה מינ' הדיוק יורד ל~65%)
2. ככל שיש יותר גיזום של העץ בC4.5 מדויק פחות והעץ בCART מדויק יותר. הסבר אפשרי הוא כי העצים מבצעים גיזום בצורה שונה :

CART מבצע גיזום ע"י מודל סיבוכיות-עלות עם פרמטרים הנגזרים מהcross validation לעומת C4.5 המבצע מעבר יחיד על העץ וגוזם לפי איבוד אינפורמציה פוטנציאלי על פי הconfidence level המוגדר.

נראה כי בC4.5 לא מתקיים overfitting בגלל הדרישה לכמות סבירה על תצפיות בעלים ולכן איבוד אינפורמציה בגיזום רק פוגע במודל.

1. חשוב להריץ את האלגוריתמים מס' פעמים עם פרמטרים שונים על מנת להבין יותר טוב את המידע איתו אנו עובדים שכן כללי אצבע לא תמיד עוזרים בבחירת פרמטרים  
   לדוגמא, נצפה כי Pruning של העץ יכול לפגוע בפרמטרים מסויימים (במקרה שלנו זה היה דווקא בפרמטר הנחוץ לנו)
2. חשוב להכיר היטב את המידע שעובדים איתו. ראשית, למדתי על סט שהכיל את שדה duration והתוצאות הדיוק היו מרשימות בהרבה מאלו המוצגות כאן. אולם בחינה נוספת העלתה כי שדה זה אינו רלוונטי למידע אמיתי שכן לא קיים עבור תצפיות לא מסווגות ובמקרים שכן קיים ישנה קשר חזק בינו לבין המשתנה התלוי ולכן אסור ללמוד עליו.

פרמטר נוסף אשר מעניין אותנו בעולם האמיתי היא כמות אבסולוטית של TP שהרי גם המודל המדויק ביותר לא עוזר לנו אם אינו אנו מסוגלים לייצר רשימת leads לטלפנים.

לאור זאת, נראה כי C4.5 המדויק יותר, עדיף על CART הן לאור אחוז הדיוק הגבוה יותר והן לאור כמות הTP הגבוה יותר בריצה ההיא.

ולכן נבחר במודל C4.5 עם פרמטרי ריצה 8.

הצעות לשיפורים :

לדעתי במקרה זה אין משמעות רבה למדדי דיוק אבסולוטיים שכן לא מאוד מעניין מה הדיוק או איכות כל קבוצת הyes שנבחרה אלא רק על subset נבחר. אפרט :

בעולם האמיתי אין אנו נדרשים לבחור בכל קבוצת הסיווג אלא בכמות מסויימת של leads לפי דרישה. נניח כי הבנק יכול לטלפן ל100,000 לקוחות בחודש ויש לו רשימת leads של10M. היינו מריצים את המודל שבנינו בפרוייקט זה על הרשימה ומתוך קבוצת הyes היינו צריכים לבחור רק את ה100k הטובים ביותר, לכן השאלה האמיתית שצריכה להישאל היא מה איכות המודל בחיזוי לקוחות שיחתמו על פיקדון בבחירת 100k אנשים. על מנת לבחור בצורה מיטבית מתוך קבוצת הyes נרצה להשתמש במודל אשר יודע להוציא score לכל רשומה בTest (אשר יכול לציין לדוגמא את ההסתברות של הרשומה להשתייך לקבוצה המסווגת) כגון CHIAD. אזי, נוכל לבחור ב100k הלקוחות עם הScore הגובה ביותר.

לאחר בחירה זו, לדעתי המדד הטוב ביותר לבחינת המודל יהיה lift על גבי random או מודל אחר אם קיים. במידה ומדובר בrandom אז נקח את אחוז הyesים מתוך הקבוצה ואז אחוז הyesים מתוך ה100k שסיווגנו, האחוז השני חלקי הראשון יהווה Lift של המודל (או – פי כמה המודל שלנו מסווג יותר טוב מבחירה רנדומלית).

ככל שהlift יהיה גבוה יותר כך המודל יהיה טוב יותר.