תרגיל 10

מודלים סטטיסטיים ב

שאלה 1

בשאלה זו נשתמש בנתונים בשם bank הקיימים כקובץ csv במודל.

אלו נתונים בנוגע לקמפיין של רשת בנקים בפורטוגל, שמטרתו הייתה לשכנע אנשים לקחת הלוואה. כל שורה מייצגת חשבון לקוח/ה והמשתנה אותו נרצה לחזות הוא האם הלקוח/ה לקח/ה הלוואה. המשתנים הנתונים הם:

- y - האם לקח/ה הלוואה (כן/לא)
- age - גיל (נומרי)
- housing - האם ללקוח/ה יש משכנתא (כן/לא)
- university - האם הלקוח/ה עשה/עשתה תואר באוניברסיטה (כן/לא)
- contact - המספר אליו נערכה השיחה (סלולרי/קווי)
- duration - אורך השיחה בשניות (נומרי)
- previous - מספר הפניות הקודמות שנעשו ללקוח/ה לפני קמפיין זה (נומרי)
- marital - מצב משפחתי (רווק/ה, נשוי/אה, גרוש/ה)

נרצה לבנות מודל רגרסיה לוגיסטית על מנת לחזות הלוואה. תוכלו להשתמש בשבעת המשתנים המסבירים הנתונים: גיל, האם ללקוח/ה יש משכנתא, האם עשה/עשתה תואר באוניברסיטה, המספר אליו נערכה השיחה, אורך השיחה בשניות, מספר הפניות הקודמות ומצב משפחתי (ללא אינטראקציות).

א. כמה מודלים אפשריים קיימים (תוכלו להשתמש בשבעת המשתנים המסבירים הנתונים, ללא אינטראקציות, עם אפקטים לינאריים בלבד למשתנים הרציפים)?

ב. השתמשו ב Backward elimination על מנת לבחור מבין המודלים לפי ה AIC שלהם.
איזה מודל נבחר כטוב ביותר?
התאימו את המודל הנבחר והסבירו את הפירוש של האומדים לפרמטרים שלו.

ג. השתמשו ב Forward selection על מנת לבחור מבין המודלים לפי ה AIC שלהם.
איזה מודל נבחר כטוב ביותר?
התאימו את המודל הנבחר והסבירו את הפירוש של האומדים לפרמטרים שלו.

ד. השתמשו ב Backward elimination וב- Forward selection יחד על מנת לבחור מבין המודלים לפי ה AIC שלהם.
איזה מודל נבחר כטוב ביותר?
התאימו את המודל הנבחר והסבירו את הפירוש של האומדים לפרמטרים שלו.

שאלה 2

נשתמש בנתונים בשם bank הקיימים כקובץ csv במודל.
נרצה לבנות מודל רגרסיה לוגיסטית החוזה הלוואה. לשם כך נרצה לבחור אחד משלושת האופציות למשתנים מסבירים:
- גיל, האם עשה/עשתה תואר באוניברסיטה ואינטראקציה שלהם.
- גיל, האם יש משכנתא ואינטראקציה שלהם.
- האם יש משכנתא, האם עשה/עשתה תואר באוניברסיטה ואינטראקציה שלהם.
השתמשו Leave One Out Cross Validation (LOOCV) על מנת לבחור מבין שלושת המודלים את המודל הטוב ביותר מבחינת Deviance.

שאלה 3

א. נשתמש בנתונים בשם bank הקיימים כקובץ csv במודל.
נרצה לבחון שני מודלים לוגיסטיים. המודל הנבחר בשאלה 1 סעיף ד', והמודל הנבחר בשאלה 2. התאימו את שני המודלים.

ב. השתמשו בפונקציה hoslem.test, מהחבילה ResourceSelection, על מנת לבצע Hosmer-Lemeshow test לכל אחד משני המודלים שהתאמתם בסעיף א.

I. עשו זאת פעם אחת עם חלוקה ל5 קבוצות. מה ניתן ללמוד מהתוצאה?

II. עשו זאת פעם נוספת עם חלוקה ל10 קבוצות. מה ניתן ללמוד מהתוצאה?