ע"י דוד פרנס, אריאל נובומינסקי ואילי ששון
ת.ז. 337977045, 32228974, 322995358

בונוס:

הגדרת המשימה: במקום להתייחס לבעיית חיזוי/רגרסיה כמו בחלק העיקרי של הפרויקט, בחרנו להסתכל על בעיה של זיהוי דפוס. זאת אומרת, רצינו בניתוח נתונים לזהות דפוסים שמובילים לתופעה ולהשתמש בכך במודל שלנו. בפרט, חלקנו את סט הנתונים שלנו לשלושה קטגוריות: יום עם כמות משקעים קיצונית, יום עם כמות משקעים רגילה ויום בלי משקעים. בהמשך נקרא לסיווג הזה משתנה היעד.

התייחסנו לזה בתור בעיה כללית ולא התייחסנו למימד הזמן מעבר לימים שקדמו ליום שמנבאים. על כן הגדרנו סט אימון וסט מבחן באופן רנדומלי על פני כל הנתונים.

כדי לבצע את המשימה העשרנו את סט הנתונים שלנו עם כמות המשקעים בימים שקדמו ליום שמנבאים. בסופו של דבר סכמת הנתונים שהמודל רץ עליו הוא:

- תחנה
- תאריך
- משקעים ביום לפני
- משקעים יומיים לפני
- משקעים שלושה ימים לפני
- חודש
- שנה
- קלאסטר

נסביר את הפיצ'רים הללו.

תחנה ותאריך מהווים מפתח של האירוע שאנחנו מנסים לזהות.

בניתוח נתונים מצאנו שקיים קשר בין המשקעים בימים הקודמים לבין המשקעים באותו יום שמנבאים. הקשר הכי חזק הוא בין היום לפני לבין יום הניבוי עם קורלציה בערך 0.15. בנוסף, מצאנו קשר בין זמן לבין משתנה היעד. בפרט, בשנים האחרונות יש יותר ימים שהם בלי גשם. וגם שבדרך כלל ימים עם כמות משקעים קיצונית קורים באמצע השנה. לכן, גם השתמשנו עם מימד הזמן בשביל זה. בנוסף, זיהינו דפוס במשקעים בימים הקודמים בעזרת KMEANS. עבור k=4 מצאנו קלאסטרינג עם silhouette של 0.711. שזה ציון גבוה. אחרי הסתכלות על הקלאסטרינג האלה נמצא שיש 4 קבוצות. 1. לא היו הרבה משקעים בימים לפני (על בסיס הצנטרואיד). 60% של המקרים האלה היו שייכים לסיווג של לא ירד גשם. הקלאסרים האחרים הם יותר קטנים בגודלם. וראוי לציין שבקלאסטר האחרון יש דפוס של ירידה בכמות המשקעים בימים הקודמים ויחד עם זה היה את הריכוז היחסי הכי גבוה של ימים שסווגו עם כמות משקעים קיצונית. לכן, אנחנו מאמינים שקלאסטרים אלו מצביעים על קשרים שלא נתפסים על ידי רק קיום הנתונים של המשקעים בימים הקודמים.

בחרנו למדוד את הביצועים שלנו ע"י f1. זאת מכיוון שרצינו למצוא מודל שאיזן בין precision וrecall. השיקול הזה מגיע מאחד מחברי הצוות שלנו. ברמה האישית הוא היה גר באזור בארה"ב אשר חווה מידי עונה תקופות של כמות משקעים קיצונית (נקרא flash flooding) וזה היה מציף את הנהר הקרוב (נהר המיזורי) והורס נכסים. לכן, מצד אחד חשוב הrecall כדי לא לפספס אזהרה שהולך להיות כמות

משקעים גדול וצריך להתכונן. מצד שני, עקב העלות של הכנות כאלה גם יש שיקול של precision לא לנבא לא נכון יותר מדי. F1 מיועד כדי למצוא איזון זה. לכן, בחרנו בו בתור המטריקה שלנו.

המודל שבחרנו בו הוא Naïve Bayesian Classification עם פרמטר Laplace Smoothing של 1. בחרנו במודל זו בגלל ההיגיון הפנימי של חשיבה באסיאנית. ראינו שאין קשרים פשוטים בין הפיצ'רים לבין משתנה היעד אך אספנו הרבה פיצ'רים שכל אחד מהם בנפרד מרמז בכיוון של משתנה היעד. כלומר, קשר של בהינתן הפיצ'ר אז יש יותר סיכוי שיהיה משתנה היעד. זה הקשר שמודל באסיאנית תופסת. אמפירית, שיפרנו את הf1 לעומת הbaseline של RandomForest ב0.17.