פרויקט האקטון של אביאל שטרן שי עזריאל רוני איטקין ואיתמר מירון

איזה אתגרים יש בדאטה?

הוא על הפנים. יש הרבה מידע חופף לדוגמא: מיקום שנתון בקווי אורך ורוחב. בנוסף השאלה ששאלנו את עצמנו היא לאיזה רזולוציות אנחנו רוצים לרדת עם הדאטה לדוגמא כמה ספציפי נרצה שיהיה המידע שנחזיק עבור הלומד עם משתנים כמו מחוז, אזור מסויים ובלוק משטרתי.

יש מידע מגוון שמוחזק בפיצר אחד - זמן ותאריך. מצד אחד התאריך מתאר עונה בשנה אבל גם יכול לתאר יום בשבוע וסביר ששניהם מגלמים בתוכם נתונים שמשפיעים על הדאטה בצורה שלא ברורה לנו.

בעיה ראשונה - זיהוי סוג פשע:

אופן הטיפול בדאטה

עמודות שזרקנו:

1. העמודה ה־0 הכילה אינדקס סידורי לפי סדר השורות ולכן לא מתארת דבר רלוונטי.
2. id לא התקבלו במדגם דגימות כפולות וראינו שהמודל שה־id לא נמצא בקורלצייה נמוכה לסוגי פשעים.
3. $"year"$ כל הדגימות מאותה שנה.
4. $"location", (x, y)$ וקווי אורך ורוחב מייצגים בטאפל את קווי האורך והרוחב שכבר יש להם עמודות משלהן, ראינו שקורלציית פירסון שלהן עם סוגי פשעים היא נמוכה מאוד.
5. $Ward$ ראינו שהוא מתואם עם $district$ אבל ש־$district$ יותר קטן ונוח לעיבוד.
6. $Block, beat$ הערכנו שזאת ירידה לרזולוציה נמוכה מידי של מיקום והעפנו כדי להימנע מאובר פיטינג (וגם כי זה גרם לשגיאת הכללה יותר גדולה הערכנו שזה בגלל רעש).

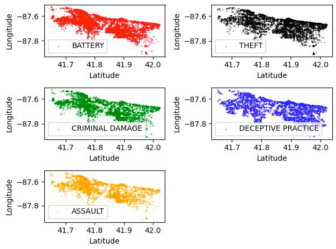
משתנים שערכנו

1. תאריך וזמן: ראינו שיש קורלציה גבוהה בין השעה לסוג הפשע ולכן זיקקנו רק את השעה מתוך הפרמטר הזה.
2. רשימת מילים בתיאור המיקום שהייתה להם קורלצייה יחסית גבוהה עם הופעת סוגי פשעים ושכיחות הופעה גבוהה.
3. עשינו דאמיז **(טיפשונים)** ל־$District$ כיוון שראינו שיש קורלציה בין שכיחות סוגי פשעים לבין המיקום והחלטנו שמחוז נותן מספיק מידע עם מספר ערכים יחודיים יחסית נמוך.
4. המרנו את כל ערכי האמת והשקר בדאטה ל־0 ו־1 בהתאמה והשארנו אותם בגלל קורלציה גבוהה.

בחירת מודל

איזה מתודות ניסינו ואיזה תוצאות הן הניבו?

תחילה ניסנו לחשוב על הבעיה בכללותה, ניסנו להסתכל במבט מלמעלה על הדאטה ויצא ששיקגו מפוצצת פשע:



ולכן מהתבוננות בדאטה הבנו שזאת בעיה שלא נכול לפתור ע"י מפריד לינארי פשוט, לכן הרצנו בלולאה knn, עץ החלטה, ויער החלטה מקרי.

כיוון נוספים שהיו לנו שהתרסקו בקלפי:

1. בניית מודל נפרד לכל סוג פשע: משם ניסנו להגיע למטריצה המתארת את החלטת כל אלגוריתם וניסנו לבנות מודל חדש שיקח את מטריצת התוצאת האפשריות שהתקבלה ויסווג אותה משם.

2. בניית עץ החלטה לכל $disctrict$: ראינו שהתפלגות סוגי הפשע על כל מחוז שונה וניסינו לבנות עץ החלטה שממיין בשלב הראשוני לפי ה$district$ ומפעיל עליו אלגוריתם של יער מקרי \ ועדת עצי החלטה \ knn שתתאר את הדאטה הצפוי.

בסוף הגענו למסקנה ע"י השוואה ערכי ACC שיער מקרי מתאר בצורה המיטבית את הבעיה שלנו מבין האופציות שחשבנו עליהן.

שגיאת הכללה מצופה: 0.5

בעיה שנייה - שליחת שלושים ניידות:

אופן הטיפול בדאטה

במקום להזכיר את המשתנים שזרקנו יהיה יותר רלוונטי לדבר על אלה שהשארנו ומדוע:

1. קורדינטות (x, y) : הבעיה המתוארת היא בראשיתה בעיה גיאומטרית ולכן לקחנו את התיאור הנ"ל למרחב כיוון שהוא היה יותר נוח מקווי רוחב ואורך והוא הפרמטר של דרישת הפלט.

2. תאריך: לקחנו את העמודה ויצרנו ממנה שתי עמודות - אחת של יום בשבוע ושנייה של זמן ביממה. זאת כיוון שהזמן הוא תיאור של מימד נוסף באופן פתרון הבעיה הגיאומטרית וראינו קורלציה בין היום בשבוע למיקום שכיח של פשעים.

בחירת מודל

כפי שכבר ציינו הבעיה המוצגת היא בעיה של שכיחות תופעה במרחב שנוח לתארו כתלת מימדי $(x, y, t) \in \mathbb{R}^3$ כאשר t מייצג זמן. הקורדינטות (x, y) חסומות ע"י הגבולות של העיר שיקגו וציר הזמן חסום ב- $[0, 24]$. זאת בעיה של $clustering$ (או לפחות, טבעי לחשוב עליה כעל כזאת) ופתרנו אותה באמצעות האלגוריתם $k-means$.

איך בחרנו את k?

הבעיה הגיאומטרית המתוארת מתארת למעשה צורה של גליל ב-$\mathbb{R}^3$, כאשר גובה הגליל הוא הזמן ורדיוס המעגל בבסיס הוא המרחק של הניידת שנשלח מכל פשע שהיא תתפוס. על מנת להגיע לקירוב ראשוני ל-k חילקנו את השטח של העיר שיקגו בשטח של מעגל ברדיוס 500 והגענו למספר בין 700 ל-800. בסוף לאחר כמה תצפיות נומריות החלטנו ש-800 הוא ה-k שממשיך איתנו לשלב הבא. כמו כן האלגוריתם מותאם לכל יום ספציפי בשבוע.

פונקציית הפסד:

בהינתן צנטרואיד המתקבל ע"י האלגוריתם $k-means$ נחשב כמה מתוך הסמפלים של פשעים המשוייכים אליו נמצאים במרחק של לכל היותר 500 מטר ממנו וחצי שעה וניקח את 30 הצנטרואידים עם המספר הגדול ביותר. בפרט פונקציית ההפסד מוגדרת היטב כל האלגוריתם $k-means$ הוא פונקציה של הסאמפלים משמע משייך כל נקודה לצנטרואיד אחד ולכן אנחנו יודעים שלא ספרנו הצלחות פעמיים.