התחלנו מלהריץ h2o על מנת לקבל תמונה כוללת ולדעת איך כל פיצ'ר משפיע על סוג הפשע(AutoML for hackathon).

לאחר מכן הסתכלנו על הדאטא וראינו שיש הרבה משתנים קטגוריים כמו BEAT WARD DISTRICT COMMUNITY AREA BLOCK

ניסינו לבדוק את הערכים ומצאנו שיש רק שני ערכים שאין בהם הגיון (NAN) ולכן הורדנו את השורות האלה (2 דוגמאות מתוך 35000, יכולנו לוותר עליהן)

כעת הסתכלנו על הפיצרים LATITUDE LONGITUDE LOCATION Y.COORDINATE X.COORDINATE

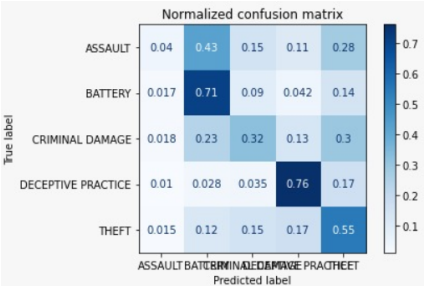
LOCATION הוא חיבור של LONG LAT ולכן אין בו טעם,כמו כן LONG LAT ו XY מסמנים לנו נקודה ספציפית, ומכיוון שיש לנו את ה BEAT שנותן לנו איזור יותר גדול השאלה היא האם יש טעם בכלל, ההחלטה היא ש פשע בנקודה מסויימת לא נותן לנו מידע אלא האזור שבו הוא התרחש (כלומר הסיכוי שפשע יתרחש פעמיים באותו מקום הוא לא גדול לעומת זאת שכונת פשע הוא משהו שקיים), ולכן הורדנו את הפיצרים הללו.

לאחר מכן הסתכלנו על הפיצרים של ה DATE וה UPDATED ON החלטנו שלימים והשעות שבו הפשע קרה יש משמעות לפשעי עתיד (לדוגמא רוב הפשעים התרחשו בימי חול , בשעות הלילה) ולכן חילקנו את התאריך לימים בחודש בשבוע ושעות ביום כמשתנה קטגורי.

החלטנו שיכול להיות יחס בין ה כמה זמן התיק מתמשך לבין חומרת הפשע ולכן החלטנו להחליף את הפיצר UPDATED ON במרחק בימים בין התאריך שבו הפשע קרא לתאריך הנל

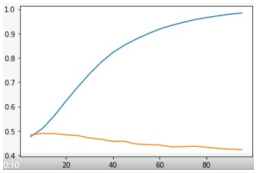
FLOW

ניסינו בהתחלה רגסיה לוגיסטית וקיבלנו בערך 50% הצלחה



לא ראינו אוברפיט (המודל נתן ביצועים דומים על הטריין ועל הולידישן) ולכן החלטנו לנסות מודל שונה.

ניסינו עצים שנעים בעומק בין 5 ל 80



אפשר לראות מהגרף שההצלחה על הטריין עולה ושואפת ל 100% לעומת זאת ההצלחה על הולידישן נשארת נמוכה (סביב ה 50%). כלומר אנחנו באוברפיט.

הבנו שהבעיה היא בהכללה וניסינו יער עצים (להקטין את האוברפיט)

ניסינו עם כל מיני פרמטרים של די קורלציה (כמה פיצרים לבחור בכל פעם) ועם עומק של 10-30

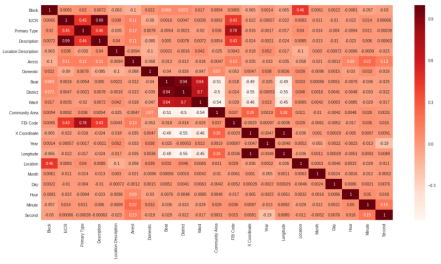
מפאת קוצר הדוח לא נתאר את כל הניסויים.

כאשר ככל שעומק העץ גדל כך ההצלחה על הטריין עולה אך את ההצלחה על הולידיישן לא הצלחנו להעלות מעל 52.5%

בשלב הזה חשבנו שאולי מכיוון שיש ריבוי פיצרים שרובם קטגורים כלומר ONE HOT וקטור הדאטא ספארסי מדי ולכן המודלים לא מצליחים להכליל, ניסינו להוריד את המימד בעזרת PCA אבל התוצאות לא השתפרו.

לאחר מכן ניסינו לעשות "הורדת מימד ידנית", כלומר לחפש פיצרים שלא תורמים לנו ולהוריד אותם, כאשר פיצרים שלא תורמים לנו הם פיצרים שההתפלגות על ה PRIMARY TYPE זהה איתם או בלעדיהם. גם זה לא עזר ולא הצלחנו להרים את ההצלחה על הולידיישן.

אנחנו מתחילים לחשוב בשלב הזה ששכונת פשע לא מתאפיינת בסוג אחד של פשיעה כלומר אם השכונה היא שכונת פשע אז כל הפשעים מתבצעים בה ולכן קשה לקלאסיפייר לדעת איזה פשע בדיוק נעשה. אולי כדאי לרדת לרזולוציה נמוכה יותר כמו BLOCK אבל כרגע אין לנו מספיק דאטא בשביל להפוך את ה BLOCK לפיצר משמעותי מכיוון שיש יותר מדי אפשרויות שונות עבור ה BLOCK (לפחות אנחנו לא מצאנו דרך להשתמש בו) .



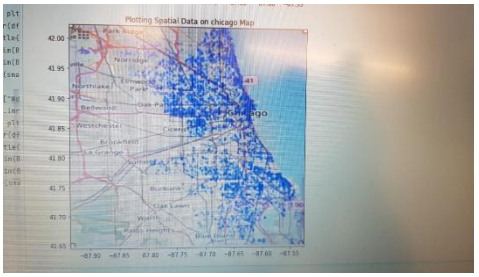
זאת מטריצת הקורלציה בין העמודות, ניתן לראות שהפיצרים היחידים שיש להם קורלציה עם הPrimary Type הם הפיצרים שבהם אסור לנו להשתמש, דבר שאולי יכול להסביר למה אנחנו לא מצליחים להסביר בעזרת הדאטא את הלייבל.

משימה 2

ניסינו וניסינו, ומה לעשות בסוף מגלים שמאחורי למידת מכונות מסתתרים בני אדם שעדיין לא למדו.

היה לה הרבה מחשבות על איך פותרים את הבעיה השנייה בעזרת unsupervised learning בסוף הלכנו על משהו שהוא בכלל לא אלגוריתם למידה קלאסי, אבל חשבנו שהוא ייתן איזשהו תוצאה, האלגוריתם לוקח את הדאטה מחלק אותו לסופי שבוע, אמצ"ש, בהתחלה רצינו ללכת על כל יום בשבוע אבל זה יותר מדי variance אז ויתרנו, אז מחלקים את הדאטה ככה לשניים, ואז מחלקים עוד פעם ל6 טווחי שעות אפשריים, אחר כך לוקחים את המפה ומחלקים אותה לריבועים של קילומטר על קילומטר, וסופרים כמה פשעים יש בכל ריבוע, וזה האימון

שלב הפענוח טוענים את הטבלה הענקית הזאת, ואז לפי התאריך בודקים אם נמצאים באמצ"ש או סופ"ש, ומתוך הריבועים שהיה בהם הכי הרבה פשיעה, לוקחים את חציון השעות חציון הX והY, וככה בעצם שולחים ניידת לאיזור שהכי מועד לפורענות בשעה שהכי מועדת לפורענות וזה נשמע שאמור לעבוד, אבל לא עובד חבל



ניסיון שני:

חילקנו את הדאטא לאמצע שבוע ולסוף שבוע ובתוך כל יום לאינטרוולים של שעות.

הנחנו שרוב הפשיעה מתרחשת סביב אותם מרכזי פשע ולכן בעזרת kmeans מצאנו את המרכזים התלת מימדיים של השעות, קו רוחב וקו גובה.

שלב האימון: חילקנו את הדאטא לפי אמצע שבוע וסוף שבוע ועל כל חלק אימנו מודל kmeans)הנחנו שיכול להיות הבדל גדול בין מרכזי הפשעים באמצע שבוע לבין מרכזי הפשעים בסוף השבוע).

שלה החיזוי - בדקנו לפי התאריך איזה יום זה בשבוע ושלחנו ניידת למרכזים שקיבלנו בשלב האימון לדאטא המתאים.

שלב הבדיקה: בחרנו תאריך ולקחנו רק את הפשעים שהתרחשו באותו תאריך. לפי התאריך חישבנו את היום שבו התרחשו הפשעים, קיבלנו מהמודל את מרכזי הפשיעה המתאימים ובדקנו את המרחק בין כל מקום בו התרחש הפשע לבין כל המוניות שהצבנו. ההצלחה שלנו נמדדה בכמות הפשעים שהצלחנו למנוע

לא הצלחנו להבין איך המרכזים חוזרים בkmean, חבל.