האקתון

תיאור הדאטה:

קיבלנו דאטה עם 13 פיצ'רים ו36,000 דגימות של פשעים בעיר שיקגו, המתארים את הפשעים לפי סוג, מיקום ברמות שונות, זמן, ומאפיינים שונים אחרים.

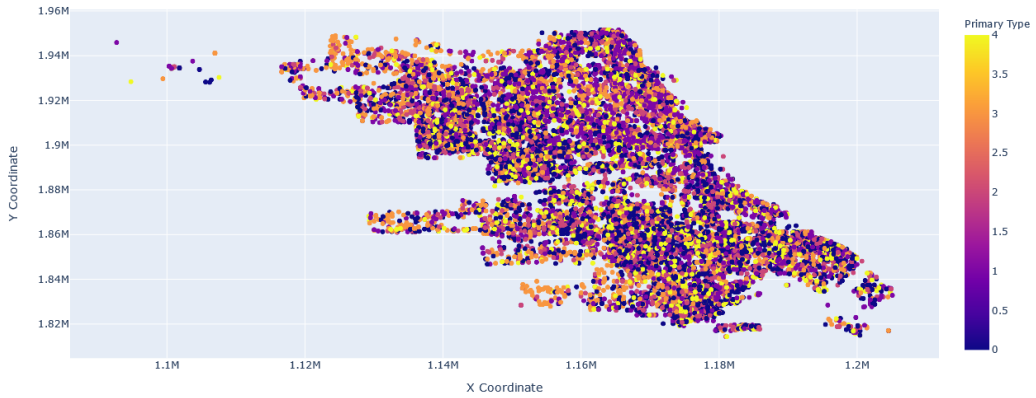
מאפיינים בעייתים בדאטה:

- מאפיינים בעייתים במשימה הראשונה: הרבה מהפיצ'רים מכילים מידע מיקומי, ומתוכם משתנים כמו Block, district, ward מכילים מידע מספרי וטקסטואלי שאין לו קורלציה ישירה למיקום במרחב.

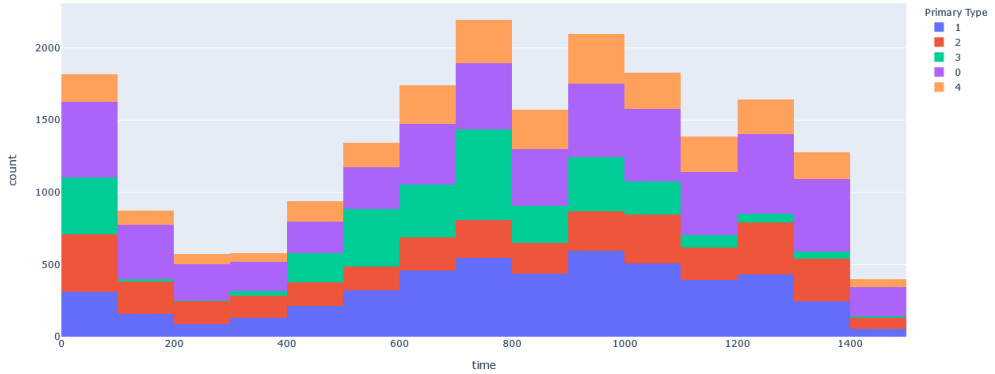
- מאפיינים בעייתים במשימה השנייה:

למרות שבשלב הקלאסטרינג רצינו להתחשב ביום בשבוע של הפשעים, נתון זה לא ניתן לנו באופן ישיר. כדי להתמודד עם בעיה זו השתמשנו בפונקציה שיודעת עפ"י בהינתן תאריך להגדיר באיזה יום בשבוע הוא היה. בנוסף היה צורך להמיר את השעות לפורמט לינארי כדי להשתמש בו בקלאסטרינג, אז להמיר אותו חזרה לפורמט של זמן. פתרנו את הבעיה ע"י המרת כל שעה למספר הדקות שעברו מאז חצות של אותו יום ועד אליה.

מיפוי של קואורדינטות X,Y עם צבע לסוג עבירה:



היסטוגרמה של עבירות במהלך היום עם חלוקה לפי סוג:



-

-

תיאור שלב ה preprocessing:

- לפני שלב הpreprocessing חילקנו את המידע שלנו לtrain,validation,test ביחס של 60-20-20

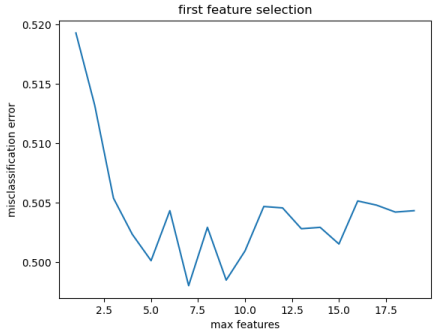
מטלה ראשונה:

מכיוון שישנן מספר עמודות המתארות את מיקום הפשע, החלטנו לוותר על מספר עמודות, ביניהן

Block. את עמודת בחרנו לפצל ליום, שעה, תאריך , ויום בשבוע.
שמנו לב שעשויה להיות קורלציה בין location description לסוג הפשע, התבוננו במיקומים הנפוצים ביותר ובחלקם אכן התפלגות הפשעים הייתה שונה מההתפלגות בדאטא הכולל. לכן יצרנו עמודות שבהן יש 1 או 0 לגבי הימצאות במיקומים נבחרים.
עבור כל עמודה, החלפנו ערכים ריקים בערכים דיפולטיבים.
<u>מטלה שניה</u>:
בחרנו להשתמש בשעה, יום בשבוע ובמיקום (X,Y).
בשלב מאוחר יותר נעשה קלאסטרינג לפי השעה והמיקום, פר יום בשבוע, כך שאלו סוגי הדאטה הרלוונטיים ביותר.

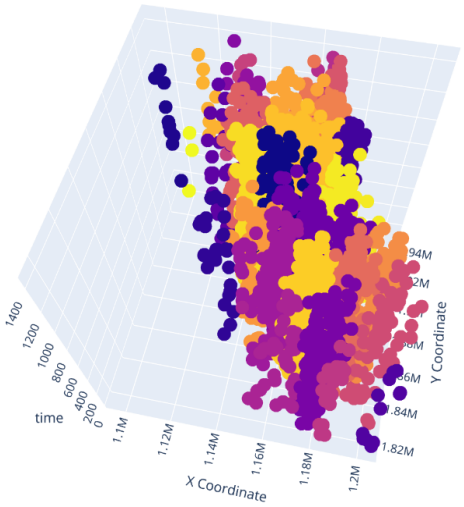
<u>שיקולים בעיצוב המודל ותיאור המודל שבסוף מימשנו:</u>

- **בחירת מודל למשימה הראשונה:** תחילה שמנו לב שקיים קשר גדול בין המיקום לבין הסוג של הפשע, לכן רצינו מודלים שמתחשבים במיקום של הנקודות וגם בשעה. ניסינו להשתמש בKNN ובסוגים שונים של עצים בניהם RandomForest וAdaboost TreeStump. השווינו בין המודלים ולאחר הערכה של המודלים על הvalidation data ראינו שrandom forest נותן את הביצועים הטובים ביותר - misclassification error הנמוך ביותר. לכן בחרנו ביער האקראי וניגשנו לבחירת ההיפר הפרמטרים שלו. לכל היפר פרמטר היינו צריכים לעבור על הערכים שלו ולראות איזה משיג את התוצאה המיטבית, לשם כך , היינו צריכים לקבע את הפרמטרים האחרים. בגרף מוצגת ההשפעה של הערכים השונים של "מספר הפיצ'רים המקסימלי בחלוקה" על ביצועי המודל



בגרף ניתן לראות שיש plateau סביב 9 ולכן בחרנו בערך הזה. כך המשכנו עם שאר ההיפר פרמטרים.
שגיאת המיסקלסיפיקציה על הTest הייתה 0.5021037868162693 ואנו צופים ששגיאת ההכללה תהיה דומה.

- **בחירת מודל למשימה השנייה**: החלטנו לממש אלגוריתם למידה לא מפוקחת של spectral clustering. בחירה זו נעשתה לאחר שזיהינו שהמרחק בין פשע לפשע מהווה חלק משמעותי מהחיזוי - הרי נרצה לשלוח ניידת לאזור צפוף בפשעים כמה שרק אפשר - אם נשלח ניידת אחת בזמן ומיקום צפוף בפשעים נוכל למקסם את מספר הפשעים שנמנע.
את פרמטר האפסילון למודל בחרנו בצורה הבאה: ראשית רצינו להגדיר לעצמינו איזה מרחק אנו מגדירים כ"קרוב" בין 2 נקודות בדאטה, כלומר איזה מרחק נרצה שיגיד ששתי נקודות יהיו באותו הקלאסטר. לאחר וויזואליזיה של הנקודות בחרנו במפה שתי נקודות קרובות וחישבנו את הנורמה האוקלידית בינהן, וכך בחרנו את אפסילון.
לאחר שהגדרנו את ההיפר-פרמטר אפסילון, חישבנו את מטריצת האפינינט והרצו את האלגוריתם, קיבלנו תיוג לקלאסטרינג לכל אחת מהנקודות בtraining data. לאחר מכן חישבנו את ה centroid של כל קלאסטר ושלחנו לשם את כל הניידות למיקסום צפיפות הפשעים.



בתמונה: Spectral Clustering of Chicago