# האקתון $Challange\ 2$ - $IML$

שמות :יובל רודיטי, אביטל קסוביץ', שיר שניארסון ובן גפרית

7 במרץ 2022

## 1 שאלה ראשונה - קלסיפיקציה

### 1.1 חקר המאגר ועיבוד מקדים

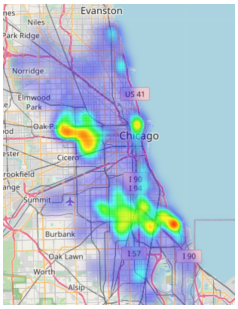
- אקספלורציה של המאגר - הבחנו במבנה הג"ג ההיררכי שיש לקובץ $District \rightarrow Community\ Area \rightarrow Ward \rightarrow Beat \rightarrow Location$, כאשר עד לרמת ה$Location$ הערכים הם קטגוריים.
- הסרנו פיצ'רים שלא יהיו רלוונטיים למודל שלנו - כמו $Case\ Number$ ו-$ID$. שדה ה-$Year$ הוא זהה עבור כל הדגימות $(2021)$ וע"כ לא רלוונטי. כך גם עבור שדה $Updated\ On$ שאינו מציין מידע הרלוונטי לפשע אלא להכנסתו ל-$DB$, וע"כ עלול להטות את המודל.
- התמודדות עם שדה יום - המרנו כל שדה לפורמט של $Datetime$, ובאמצעותו פיצלנו את היום למספר היום בשבוע $(0-6)$. בכך אפשרנו למודל לתת משקל שונה לימים שונים. פעלנו באופן דומה עעבור החודשים.
- התמודדות עם השעה - נעזרנו בחישוב מוכר של $\sin, \cos$ של השעה ביום ע"מ ליצור פיצ'ר שיהיה מחזורי (כלומר שהשעות $24$ ו-$1$ יהיו קרובים יותר בינהם עבור המודל).
- התמודדות עם פיצ'רים קטגוריים - עבור הפיצ'רים $District, Community\ Area, Ward, Beat$ נעזרנו ב-$One\ Hot\ Encoding$ לשם פיצולם.

### 1.2 יצירת פיצ'רים נוספים

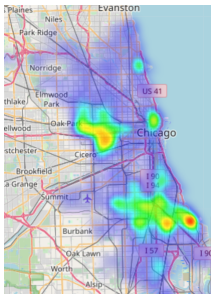
לאחר שימוש במפות חום שחלקנו לפי סוג הפשע, ראינו קורלציה חזקה מאוד בין סוג הפשע למיקום בעיר. למשל $Deciptive\ Practice$ נפוץ בעיקר במרכז העיר, בעוד $Assault$ ו-$Battery$ נמצאים מתנהגים די דומה כמצופה, ושניהם נפוצים בעיקר באיזור הפרוורים. מצד שני, התבוננות בשדה ה-$(x, y)$ לבד מייצר $Variance$ גבוה ולא עובר הכללה.
לשם כך הוספנו עוד 5 עמודות שמציינות כמה מתוך $k$ השכנים הקרובים של כל פשע היו מכל אחד מ-$5$ הסוגים. $k$ נקבע לאחר בדיקה להיות $30$.
כמו כן, לשם יעילות במימוש האלגוריתם שבודק את המרחק מהשכנים, בדקנו רק עבור שכנים שנמצאים באותו $Community\ Area$.

### 1.3 העשרה

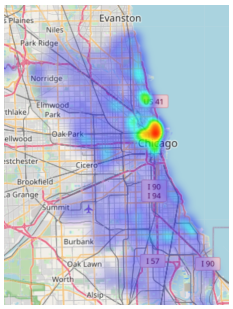
הסתמכנו על מסמך אשר כולל בו את אחוז הצפיפות עבור כל בית, אחוז האנשים שסיימו תיכון ויש להם תעודה כלשהי מעל גיל 25, אחוזי הכנסה לראש.
כאשר החלוקה הגאוגרפית נעשית בעזרת $Community\ Area$. עשינו $join$ לדאטא שלנו, שהרי גם בדאטא שלנו יש עמודה בשם $CommunityArea$, לבסוף עשינו הצמדה.
המידע הנ"ל עוזר לנו לסווג את האיזורים לאיזור מגורים, שכונות מצוקה ושכונות פחות יוקרתיות.
מידע זה עוזר לנו לאבחן ולסווג את סוגי הפשע השונים בעיר.
כלומר סביר להניח שאחוזי הגניבה יהיו נמוכים בשכונת מצוקה, בניגוד לסיכוי בחבלה.



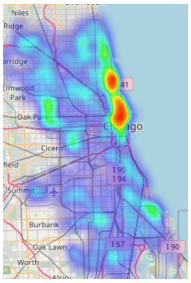

$Battery$



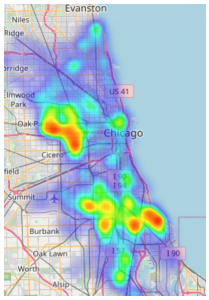

$Assault$




$Theft$




$Deceptive\ Practice$




$Criminal\ Damage$

### 1.4 בחירת מודל

בדקנו לאורך העבודה שלנו את הביצועים על חבילה גדולה של מסווגים בעזרת המודול $LazyPredict$, בהם $RandomForestClassifier$, $SGDClassifier$, $AdaBoostClassifier$ ועוד רבים. בחירת המודל התבצעה על סמך הביצועים שלו - דיוק ו$ROC\ AUC$.

ניכר כי באופן עקבי $SGDClassifier$ היה בעל הביצועים הטובים ביותר, והוא המודל שבחרנו עבור החיזוי.

| Model | | | | | |
|---|---|---|---|---|---|
| LGBMClassifier | 0.53 | 0.47 | None | 0.50 | 4.25 |
| RandomForestClassifier | 0.53 | 0.47 | None | 0.50 | 3.65 |
| XGBClassifier | 0.53 | 0.47 | None | 0.50 | 27.56 |
| ExtraTreesClassifier | 0.52 | 0.46 | None | 0.50 | 4.45 |
| AdaBoostClassifier | 0.53 | 0.46 | None | 0.49 | 5.93 |
| RidgeClassifierCV | 0.51 | 0.45 | None | 0.48 | 1.04 |
| LinearDiscriminantAnalysis | 0.50 | 0.45 | None | 0.49 | 1.22 |
| LinearSVC | 0.51 | 0.45 | None | 0.48 | 32.50 |
| RidgeClassifier | 0.51 | 0.45 | None | 0.48 | 0.33 |
| BaggingClassifier | 0.51 | 0.45 | None | 0.48 | 3.13 |
| LogisticRegression | 0.50 | 0.45 | None | 0.49 | 1.43 |
| CalibratedClassifierCV | 0.51 | 0.44 | None | 0.45 | 125.88 |
| NuSVC | 0.49 | 0.44 | None | 0.48 | 41.30 |
| DecisionTreeClassifier | 0.47 | 0.43 | None | 0.46 | 0.73 |
| BernoulliNB | 0.48 | 0.43 | None | 0.47 | 0.30 |
| SVC | 0.48 | 0.42 | None | 0.45 | 50.26 |
| SGDClassifier | 0.45 | 0.41 | None | 0.45 | 4.50 |
| NearestCentroid | 0.44 | 0.41 | None | 0.45 | 0.25 |
| ExtraTreeClassifier | 0.44 | 0.41 | None | 0.44 | 0.30 |
| Perceptron | 0.44 | 0.40 | None | 0.43 | 0.62 |
| PassiveAggressiveClassifier | 0.42 | 0.38 | None | 0.42 | 0.97 |
| LabelSpreading | 0.36 | 0.32 | None | 0.35 | 3.88 |
| LabelPropagation | 0.36 | 0.32 | None | 0.35 | 3.13 |
| KNeighborsClassifier | 0.36 | 0.32 | None | 0.34 | 5.40 |
| GaussianNB | 0.22 | 0.29 | None | 0.17 | 0.35 |
| QuadraticDiscriminantAnalysis | 0.31 | 0.29 | None | 0.28 | 2.22 |
| DummyClassifier | 0.21 | 0.20 | None | 0.21 | 0.23 |

## 2 שאלה שנייה - חיזוי פשעים

בשאלה הזאת נדרשנו לבעיית $Clustering$ גיאוגרפי באמצעות $Unsupervised\ Learning$. לצורך המשימה התסכלנו רק על שורות הפיצ'רים של מיקום וזמן. ראינו שלשעה ביום יש משקל בהיווצרות ה-$Cluster$ים, וע"כ נדרשנו ליצור $Cluster$ים שיתבססו גם על מיקום וגם על השעה ביום. נעזרנו לשם כך באלגוריתם $k-means\ clustering$ עם פרמטר $k=30$.

התאמנו לכל יום בשבוע מודל $K-means$ משלו, והחיזוי נעשה עבור היום הספציפי.

לאחר בדיקות, ראינו שהמודל מצליח לחזות בממוצע כ9 פשעים.