פרויקט בבינה מלאכותית

236502

חיזוי תאונות דרכים על סמך נתונים דמוגרפיים וגיאוגרפיים

מגישים:

רון פז | תמיר נאמן

הפקולטה למדעי המחשב, טכניון

תוכן עניינים

מבוא ותיאור הבעיה

תאונות דרכים זוהי בעיה כלל עולמית שמנסים להתמודד איתה באופן משמעותי בשנים האחרונות. בשנת 2018 נהרגו בבריטניה כ-1700 בני אדם, נפצעו בצורה חמורה כ-25,000 ונפצעו קל כ-133,000.[1]

כאשר רשות ממשלתית או מקומית מעוניינת לבצע פיתוח אזורי כגון שינוי תצורת הכבישים, פתיחת עסקים מקומיים (כמו ברים או מוקדי תרבות), או שינוי תקציב הבריאות- ההשפעה על מספר תאונות הדרכים נצפית רק כלאחר מעשה, כאשר הנזק נגרם ויש להתמודד עם השלכותיו.

לדוגמא, פתיחת בתי מרזח רבים במקום מסוים ללא שיפור תשתיות שונות עשויה להביא לעלייה חדה במספר תאונות הדרכים[2], כאשר בהינתן הידע הזה ניתן לתגבר מראש את תשתיות ההצלה והאכיפה על מנת לצמצם את מספר הנפגעים.

מהצד השני, רשויות המנסות אקטיבית להפחית במספר תאונות הדרכים יכולות היום לנסות שיטות שונות ולחכות לקבלת תוצאה בשטח, ללא הערכה מספרית לשינוי במספר התאונות שיובילו הפיתוחים השונים. מכאן שלא ניתן לבצע תיעדוף שקול בהתחשב בעלות ובזמני בנייה ושיפור תשתיות.

למשל, נחקר הקשר בין החלפת צמתים מסוימים בכיכרות לירידה במספר תאונות הדרכים[3], ואולם נותרה השאלה היכן כדאי למקם את הכיכר על מנת להרוויח כמה שיותר ירידה במספר תאונות הדרכים, או האם עדיף למשל להשקיע את העלות בפיתוח בתחומים אחרים כמו בתי ספר וגישה להשכלה שיכולים להיות קשורים בצורה עקיפה לכמות תאונות הדרכים באזורים מסוימים.

בפרויקט שלנו החלטנו להשתמש בטכנולוגיות Machine Learning ו-Deep Learning על מנת לאפשר למקבלי ההחלטות בגופים אסטרטגיים בבריטניה כמו הממשלה, משטרה וארגונים שעיסוקם בתאונות דרכים לקבל תמונת מצב אסטרטגית על תאונות הדרכים באזורים שונים במדינה, ולבחון כיצד פיתוח אזורי עשוי להשפיע על מספר תאונות הדרכים שצפויות להתרחש.

[1] "Reported road casualties Great Britain, main results: 2018 https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-main-results-2018.
[2] "Alcohol outlet density is related to police events and motor ...." 8 Nov. 2012, https://www.ncbi.nlm.nih.gov/pubmed/23216494.
[3] "Road safety effects of roundabouts in Flanders. - NCBI." https://www.ncbi.nlm.nih.gov/pubmed/16045933.

דרך לפתרון הבעיה

החלטנו לפתור את הבעיה הנ"ל באמצעות אימון מודלי Machine Learning אשר יאומנו על תכונות רלוונטיות הנוגעות לתאונות דרכים וניתנות לשינוי ע"י פיתוח איזורי, מתוך כוונה כי המודל ילמד קשר אמיתי בין אותן תכונות ומספר תאונות הדרכים במקום ובכך יאפשר לרשויות לבצע שינויים בתכונות הנ"ל (לבדוק את השפעה של שינוי מספר בתי ספר, למשל) לקבלת מספר תאונות חזוי, ולהתחשב בתוצאה המתקבלת בהחלטתם.

על מנת לעשות זאת בחרנו לחלק את המפה לאזורים גיאוגרפיים. לאחר מכן אנו מעוניינים לדלות מידע על כל אזור כמו צפיפות אוכלוסייה, כמות פאבים, כמות בתי ספר, חוסר גישה להשכלה גבוהה, רמת פשיעה ועוד. אמנם חלק מהתכונות שהצגנו כאן לא נראות רלוונטיות באופן ישיר לתאונות דרכים, אך חלק מפתרון הבעיה הוא מציאת קשרים שלאו דווקא נראים הגיוניים במבט ראשוני אך בעזרת אלגוריתמי למידה וניתוחים סטטיסטיים ניתן למצוא את הקשרים הנ"ל.

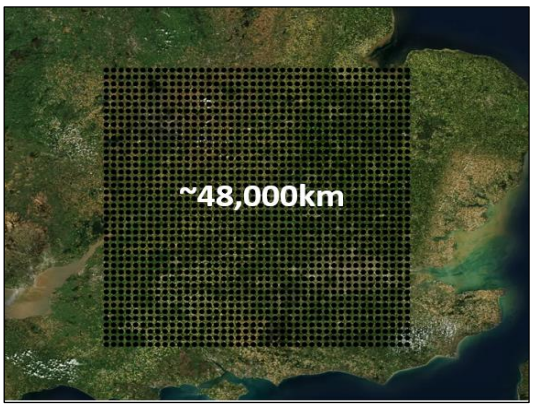
בנוסף, אנו נבדוק את הקשר בין הצורה הויזואלית של האזור הגאוגרפי הנבחר לבין רמת הסיכון לתאונות דרכים באמצעות שימוש ב-Deep Learning. אנו מעוניינים לבדוק האם בנוסף לתכונות הרגילות, תמונה של האזור במפה יכולה להוסיף לנו מידע ומאפיינים לגבי רמת הסיכון לתאונות דרכים. באופן זה אנו נאפשר למשתמש לבדוק איך שינוי במפת הכבישים עלול להשפיע על מספר התאונות באזור מסוים. למשל, כיצד החלפת צמתים בכיכרות על ידי משתמש המערכת תשפיע על כמות התאונות החזויה.

במהלך העבודה נשתמש בכמות גדולה מאוד של מידע ולכן נייצר ויזואליזציות אשר יעזרו לנו לקבל תמונות על ברורות יותר. במהלך הדו"ח נציג חלק מהויזואליזציות, אנו מאמינים שזה נותן יותר אינטואיציה ומכניס רובד נוסף של עניין בפרויקט.

חלוקת המפה למשבצות

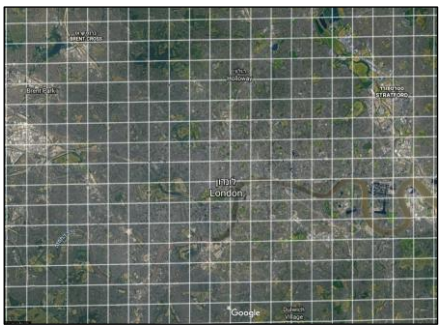
בהינתן מפה, במקרה שלנו של בריטניה, ראשית אנו נדרשים לבחור את האזור הכללי שבו נבצע את הפרויקט. בחרנו אזור בעל שטח של 48,000 קמ"ר בקירוב. לא בחרנו שטח גדול מזה מכיוון שזהו השטח הגדול ביותר שאפשר לנו הוצאת מידע משרתי מקורות המידע השונים שהשתמשנו בהם (נסביר על כך בחלק של מקורות המידע). לשם קבלת סדר גודל, שטח זה גדול מפעמיים שטחה של מדינת ישראל ולכן הוא מספיק גדול עבורנו. בנוסף, הפרויקט הוא סקיילבילי, כלומר ניתן לבחור אזור בגודל אחר ומיקום אחר ולהסיק את המסקנות בהתאם לאזור זה.

ניתן לראות ב-2 התמונות הבאות את האזור הנבחר:



תמונות על המראות את האזור הגאוגרפי שעליו עבדנו בפרויקט. נבנה בעזרת Google Earth.

לאחר מכן אנו מחלקים את האזור הכללי למשבצות בגודל $1km\ x\ 1km$. בתמונות הבאות ניתן לראות את החלוקה האמיתית שביצענו למשבצות בדו מימד ובתלת מימד:



תמונות המראות את חלוקת המפה לרשת משבצות. נבנה בעזרת Google Earth.

את חלוקת המשבצות ביצענו על ידי כתיבת אלגוריתם המקבל קואורדינטה ימנית עליונה, קואורדינטה שמאלית תחתונה (של המרובע הגדול) ומחלק את השטח שבין הקואורדינטות לריבועים קטנים בגודל $1km\ x\ 1km$. את תוצאת האלגוריתם ניתן לראות ב-2 התמונות למעלה. הויזואליזציה בוצעה באמצעות ה-grid שיצרנו והכנסנו ל-Google Earth[4]. הבחירה במשבצות בגודל $1km\ x\ 1km$ נובעת מפשטות השימוש ביחידת מידה סטנדרטית עבור משתמשי המערכת, וכן הרזולוציה המתקבלת מאפשרת לתת חיזוי ברמת השכונה הבודדת, לעומת מידות גדולות יותר אשר היו נעשות כלליות יותר, או מידות קטנות אשר היו דורשות כוח חישובי רב ולא היו מאפשרות תכנון עירוני מעבר לרמת הרחוב הבודד.

[4] "Explore Google Earth.." https://earth.google.com/web

איסוף הנתונים לפי משבצות

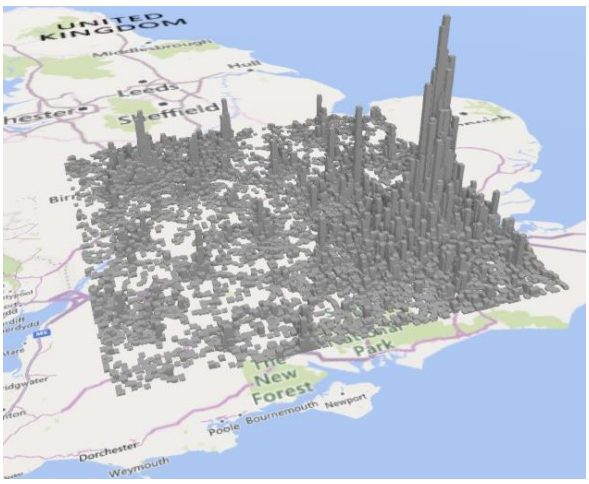
הרזולוציה שמעתה אנו מסתכלים עליה היא משבצות. לכן עלינו להתחשב בכך באיסוף הנתונים. כל שורה בסט הנתונים מציינת משבצת. לכל משבצת אספנו את התכונות האופייניות לה. כמו שציינו בהגדרת הבעיה, אספנו תכונות שונות לכל משבצת, גם כאלה שנראו לנו קשורות באופן אינטואיטיבי וגם כאלו שלא על מנת לנסות לגלות קשרים מעניינים. בין התכונות שאספנו: כמות בתי ספר במשבצת, כמות פאבים במשבצת, כמות מקומות שמתעסקים באומנות במשבצת, צפיפות אוכלוסייה במשבצת, רמת חוסר גישה לתעסוקה במשבצת, רמת חוסר גישה להשכלה גבוהה במשבצת ועוד (נפרט בהרחבה בפרק פירוט התכונות).

בנוסף לנתונים הרגילים נדרשנו גם לאסוף צילומי תמונות סטטיים של המשבצות על מנת שנוכל להשתמש בהם ב-Deep Learning. התמונות הבאות הן דוגמאות לתמונות שאספנו מתוך 48,000 תמונות (פירוט על מקור האיסוף במקורות מידע):

דוגמאות לצילומי תמונות מפה לפי משבצות שהוצאנו מ-HERE

משתנה החיזוי וסוג החיזוי עליו עבדנו בפרויקט

בחרנו במספר התאונות במשבצת כיעד לחיזוי. מספר התאונות החזוי יכול לשקף באופן מאוד ברור את רמת הסיכון למשבצת. לאחר איסוף הנתונים מצאנו שבסט האימון מספר התאונות במשבצת נע בין 0 תאונות ל-151 תאונות, בממוצע יש 1.344 תאונות במשבצת ובסה"כ יש 35,736 תאונות בתחום הנבחר.



היסטוגרמה שבנינו בעזרת פלטפורמת HERE על גבי מפה של Bing המראה את כמות התאונות לפי אזור גאוגרפי בתוך התחום הנבחר.

בחרנו לבצע רגרסיה בשיטות של Machine Learning ו-Deep Learning על מנת לקבל את מספר התאונות הצפוי במשבצת. נפרט על כך בהרחבה בשלב הניסויים.

ישנה אפשרות גם לתאר את הבעיה כבעיית קלסיפיקציה אך זאת נראתה לנו בעיה יותר קלה ופחות מעניינת (למשל לסווג למספר תאונות נמוך/בינוני/גבוה) ולכן בחרנו לפתור את הבעיה כבעיית רגרסיה.

אופן חלוקת המידע

תחילה חילקנו את ה-data לסט אימון, ולידציה ומבחן ביחס של 0.2:0.2:0.6 בהתאמה.

את שלב הלמידה ביצענו על סט האימון תוך שימוש בסט הולידציה לצורך הערכת ביצועי המודלים בזמן הלמידה.

יש לציין שסט המבחן בודד משך כל שלבי הלמידה, הוא שומש רק לאחר שלב הלמידה לצורך חישוב המדדים והערכת החיזוי הסופי.

שלב הניסויים

בטרום הלמידה ביצענו imputing ו-scaling ל-data כפי שנפרט בהמשך.

לאחר שלבים אלה המידע היה מוכן לצורך ביצוע הניסויים הבאים:

1. ניסויים לבחירת תכונות
2. ניסויים לחיזוי מספר תאונות דרכים באמצעות אלגוריתמי למידת מכונה
3. ניסויים לחיזוי מספר תאונת דרכים באמצעות למידה עמוקה
4. ניסויים לבדיקת השפעת נתוני תאונות עבר על חיזוי מספר תאונות הדרכים

מקורות מידע

החלק הראשון שבנינו במערכת היה מודולי איסוף המידע. חלק זה כלל עבודת מחקר נרחבת על מקורות מידע שונים בכדי שיתאימו לדרישות הפרויקט שלנו, הן מבחינת התוכן והן מבחינת כמות הבקשות הגדולה שלנו לשרתי המאגרים.

באינטרנט קיים המון מידע הקשור באופן ישיר ועקיף לתאונות דרכים, לכן בחרנו למקד את איסוף המידע במספר מקורות עיקריים. בנוסף, רבים ממקורות המידע דרשו תשלום במידה ועוברים סף בקשות משרתיהם. לכן, היה לנו חשוב למצוא מקורות מידע המאפשרים מספר גדול של בקשות לשרתיהם באופן חופשי, או עם הגבלה מאוד גדולה.

הוצאת המידע בוצעה בצורה מקבילית מהמקורות השונים על מנת לחסוך בזמן. לאחר מכן ביצענו הצלבות ואיחודים על מנת לקבל את מאגר המידע השלם.

החשיבות במידע מהימן ומדויק הייתה קריטית לצורך בניית מודלי הלמידה והחיזוי, לכן בחרנו במקורות מידע אמינים ומוכרים.

מקורות המידע שבהם השתמשנו:

1. **מאגר תאונות הדרכים בבריטניה**[5]- השתמשנו במאגר מידע של ממשלת בריטניה אשר מפרסמת מידע על תאונות הדרכים שהתרחשו במדינה לפי קאורדינטות. המידע כולל בין היתר מיקום תאונה מדויק על פי קוארדינטות, מספר נפגעים, חומרת התאונה, מספר כלי רכב מעורבים, LSOA (פירוט בסעיף 6) ונתונים נוספים.
2. **Foursquare**[6] - נקודות עניין- השתמשנו בפלטפורמה Foursquare על מנת לשלוף מידע על נקודות עניין בתוך משבצת במפה המוגדרת על ידי 2 קאורדינטות (פירוט בהמשך).
   Foursquare מספקים Places API שבעזרתו שלפנו את הנתונים משרתיהם. הם מאפשרים בחשבון חינמי עד 95,000 בקשות מהשרת ביממה.
   ביצענו בקרוב 285,000 בקשות להוצאת מידע ממקור זה במהלך מספר ימים.
   דוגמא לנקודת עניין שהוצאנו מפלטפורמה זו: הוצאה של כל בתי הספר הקיימים בכל משבצת ברשת המשבצות שבנינו וסכימה שלהם, לקבלת את כמות בתי הספר במשבצת.
3. **HERE**[7] - תמונות סטטיות- השתמשנו בפלטפורמה HERE לשם הוצאת תמונות סטטיות של משבצות במפה על פי קאורדינטות. בתמונות אלה השתמשנו במודל הלמידה העמוקה שבנינו כפי שנפרט בהמשך. את הוצאת המידע ביצענו משרתי HERE באמצעות Map Image API שהם מספקים. בחשבון החינמי הם מאפשרים הורדה של עד $2.5GB$. אנחנו הסתפקנו בכ- $600MB$ של תמונות, סדר גודל של 48,000 תמונות.

[5] Road Safety Data - https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data.
[6] "Foursquare Developer." https://developer.foursquare.com/.
[7] "HERE Developer - HERE Technologies." https://developer.here.com/.

4. **Google Earth**[8] - אפליקציית מפות- נעזרנו ב-Google Earth במהלך הפרויקט על מנת לבדוק את הקוארדינטות והמיקומים שעליהם עבדנו. בנוסף לזאת השתמשנו בתמונות מ-Google Earth להמחשות בדו"ח זה. רצינו לציין שבתחילת העבודה תכננו להשתמש ב-APIs שונים של גוגל מפות על מנת להוציא מידע, אך גילינו שמעבר לכמות בקשות די קטנה לשרתיהם הם דורשים תשלום ולכן השתמשנו באלטרנטיבות המצוינות בפרק זה.

5. **LSOA - Lower Layer Super Output Areas**

קודי ושמות LSOA הינם מזהים גאוגרפיים אשר הוגדרו לצורך חלוקת אזור אנגליה וויילס לתחומים קטנים לניתוח סטטיסטי, אשר חולקו בשנת 2011 לאחר מפקד אוכלוסין כך שבכל תחום LSOA התגוררו בממוצע 1500 איש, ולכל הפחות 1000. בחרנו לשלוף מידע סטטיסטי עבור מקומות שונים במפה תוך שימוש בקודי LSOA, באמצעות אוסף של מספר מאגרי מידע:

- מאגר מידע ממשלתי של הלשכה הלאומית לסטטיסטיקה בבריטניה על כמות וצפיפות האוכלוסיה[9] .
- מאגר מידע ממשלתי של-
Ministry of Housing, Communities & Local Government[10] המכיל אוסף של ציונים אשר חולקו במדדי Deprivation שונים, בחלוקה לLSOA השונים- נוכחות של פשיעה, חוסר גישה לתעסוקה, חוסר גישה להשכלה גבוהה וצבירת מיומנות, חוסר גישה לשירותי רפואה, חסמים בגישה לדיור, ולסביבת מגורים בריאה. מן הנתונים הללו אנו מצפים לקשר בין אופי האוכלוסייה והחיים בה, למספר תאונות הדרכים המתרחשות בסביבתה.

6. **Postal codes**[11] - השתמשנו ב-REST API של postcodes (פרויקט קוד פתוח) על מנת לקבל את המיקודים של משבצות שבהן לא התרחשו תאונות. במשבצות שנמצאות באזורים אלה, מכיוון שלא היו תאונות, לא היה נתון לנו מספר ה-LSOA של אותו האזור כי במאגר התאונות הממשלתי שהשתמשנו בו היה קיים מספר LSOA רק עבור מיקומים של התאונות. לשם כך הוצאנו את ה-post code ובעזרתו גזרנו מה ה-LSOA של אותה משבצת. כלומר, השתמשנו במאגר נוסף זה על מנת לבצע imputing ל-LSOA וכל הנתונים הדמוגרפיים הנגזרים ממנו.

---

[8] "Explore Google Earth.." https://earth.google.com/web.
[9] Lower layer Super Output Area population density (National .... https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareapopulationdensity.
[10] "A slice from a data cube dataset - Open Data Communities." https://opendatacommunities.org/slice?dataset=http%3A%2F%2Fopendatacommunities.org%2Fdata%2Fsocietal-wellbeing%2Fimd%2Findices&http%3A%2F%2Fpurl.org%2Flinked-data%2Fcube%23measureType=http%3A%2F%2Fopendatacommunities.org%2Fdef%2Fontology%2Fcommunities%2Fsocietal_wellbeing%2Fimd%2FdecObs.
[11] "Postcodes.io." https://postcodes.io/.

פירוט התכונות

בפרק זה נפרט על התכונות שבהן בחרנו להשתמש בפרויקט.

מספר בתי הספר (school) - תכונה המייצגת את מספר בתי הספר הנמצאים במסגרת המשבצת. אנו מעריכים כי ייתכן ומספר בתי ספר יכול להשפיע על מספר תאונות הדרכים במקום, בין על ידי עדות על אופי האוכלוסייה המתגוררת במקום (משפחות עם ילדים) או בהשפעת הילדים עצמם על מספר התאונות (התפרצויות לכביש, ראות נמוכה מאחורי מכוניות עומדות).

מספר הברים (bars) - תכונה המייצגת את מספר הברים הנמצאים במסגרת המשבצת. כפי שרשמנו במבוא, קיים קשר בין מספר מרובה של ברים למספר תאונות הדרכים במקום. אנו משערים כי הדבר נובע מהגדלת הסבירות לנהיגה בגילופין, וכן השפעה על תנועה לא צפויה של הולכי רגל בהשפעת אלכוהול.

מספר מוקדי אמנות ובידור (arts_and_entertainment) - תכונה המייצגת את מספר מוקדי האמנות והבידור במסגרת משבצת, כדוגמת מוזיאונים, אצטדיוני ספורט והיכלי מופעים. אנו משערים כי בקרבת מוקדי עניין שכאלה יופיעו לעיתים קרובות ריכוזי אוכלוסיית הולכי רגל גדולים, וכמו כן גם גדילה בגודש התעבורה. על כן אנו צופים שלמספר מוקדי האמנות והבידור תהיה השפעה על מספר תאונות הדרכים בקרבתם.

כמות אוכלוסייה (population) - תכונה המייצגת את מספר האנשים המתגוררים באזור הLSOA לו שייכת המשבצת. אנו משערים כי גודל האוכלוסייה באזור בצירוף תכונות שונות (למשל תכונות מספרי נקודות העניין) יכול להוסיף מידע אשר יאפשר חיזוי טוב יותר של תאונות דרכים. לדוגמא, נוכחות 5 ברים באזור בו גרים מאה אלף אנשים, לעומת השפעת נוכחות 5 ברים באזור בו גרים 20 אנשים.

צפיפות אוכלוסייה (population_density) - תכונה מספרית המייצגת את מספר האנשים המתגוררים בק"מ ריבועי במסגרת אזור הLSOA לו שייכת המשבצת. לצפיפות האוכלוסייה באזור יכולה להיות השפעה על מספר תאונות הדרכים על ידי השפעה על מספר כלי הרכב, וכן הולכי הרגל באזור מסוים.

ציון שליטת פשיעה (crime_dom) - ציון Deprivation שניתן על ידי ה Ministry of Housing, Communities & Local Government המתאר את דומיננטיות הפשע באזור הLSOA לו שייכת המשבצת. ערך גבוה יותר בתכונה זו מראה דומיננטיות פשיעה גבוהה יותר באזור. אנו משערים שתכונה זו עשויה להשפיע על אופי ומזג האוכלוסייה המקומית, וכן על נוכחות משטרה באזור- באופן שעשוי להשפיע על מספר תאונות הדרכים.

ציון חוסר גישה להשכלה (education_dep) - ציון Deprivation שניתן על ידי ה Ministry of Housing, Communities & Local Government המתאר את חוסר הגישה להשכלה באזור הLSOA לו שייכת המשבצת. ערך גבוה יותר בתכונה זו מתאר חוסר גישה משמעותי יותר להשכלה באזור. יתכן כי השפעת חוסר הגישה להשכלה, וכך על השכלת תושבי המקום, תשפיע על מספר תאונות הדרכים.

ציון חוסר גישה לשירותי בריאות (health_dep) - ציון Deprivation שניתן על ידי ה Ministry of Housing, Communities & Local Government המתאר את חוסר הגישה לשירותי רפואה באזור הLSOA לו שייכת המשבצת. ערך גבוה יותר בתכונה זו מתאר חוסר גישה משמעותי יותר לשירותי רפואה באזור. אנו משערים שתכונה זו עשויה להעיד על מידת ההשקעה ממשלתית באזור וכן על מזג התושבים. אנו נצרף תכונה זו מתוך רעיון כי יתכן שבאופן לא ישיר, תכונה זו משפיעה על מספר התאונות באופן שיוסיף מידע למודלי הלמידה.

ציון חוסר גישה להכנסה (income_dep) - ציון Deprivation שניתן על ידי ה Ministry of Housing, Communities & Local Government המתאר את חוסר הגישה למקור הכנסה באזור הLSOA לו שייכת המשבצת. ערך גבוה יותר בתכונה זו מתאר חוסר גישה משמעותי יותר להכנסה באזור. אנו מניחים שמקומות בהם קיים מחסור משמעותי בגישה להכנסה יהיו מקבצים של אוכלוסיות עניות יותר ומתוסכלות ממצבן הכלכלי, וכן חוסר גישה להכנסה בכל השכבות הסוציו-אקונומיות עשוי להוות גורם שיתרום למזג שלילי שיוביל לעלייה בתאונות הדרכים.

ציון חוסר גישה לתעסוקה (employment_dep) - ציון Deprivation שניתן על ידי ה Ministry of Housing, Communities & Local Government המתאר את חוסר הגישה לתעסוקה באזור ה- LSOA לו שייכת המשבצת. ערך גבוה יותר בתכונה זו מתאר חוסר גישה משמעותי יותר לתעסוקה באזור. קיימת היתכנות כי מחסור בתעסוקה ישפיע באופן מסוים על מספר תאונות הדרכים במקום מסוים עקב השפעה על מזג האוכלוסייה, ומספר השעות בהן התושבים מבלים מחוץ למתחמים סגורים.

ציון חוסר גישה לסביבת מחייה ומגורים נוחים (living_environment_dep) - ציון Deprivation שניתן על ידי הMinistry of Housing, Communities & Local Government המתאר את חוסר הגישה לסביבת מחייה ומגורים נוחים באזור הLSOA לו שייכת המשבצת. ערך גבוה יותר בתכונה זו מתאר חוסר גישה למגורים באזור. יתכן כי במקומות בהן לאוכלוסייה אין גישה למגורים תהיה השפעה על מספר הולכי הרגל שהתנהגותם בלתי צפויה ברחוב, ומידע זה עשוי לעזור בחיזוי מספר תאונות הדרכים.

תמונות המפה - עבור ניסויי Deep Learning - כפי שהוצג במבוא, קיימת השפעה של מבנה מפת הכבישים על מספר התאונות - למשל במקרה של נוכחות כיכרות בהשוואה לצמתים במפגשי כבישים. כמו כן יתכן כי רוחב כבישים, צפיפות רחובות וכן צורות כבישים בלתי רגילות

עשויות להשפיע על הסבירות לתאונות דרכים, וכך בצירוף עם תכונות אחרות - אנו צופים כי ניתן ללמוד ממבנה הכבישים על מספר תאונות הדרכים. מלבד זאת, בצילומי מפות מופיעים לעיתים אלמנטים סביבתיים כמו נוכחות של יערות, סימנים המעידים על נוכחות כבישים מהירים ועוד. אנו נשתמש בתמונות מפה על מנת לנסות לשפר ביצועיי חיזוי עבור מודל רשת נוירונים, אשר יוצג בהרחבה בפרק הניסויים.

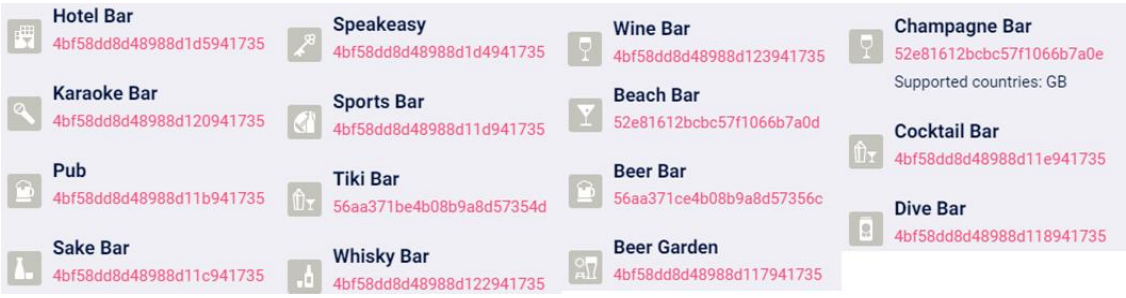
הצגת סטטיסטיקות גולמיות על הנתונים

על מנת להכיר טוב יותר את הנתונים הוצאנו היסטוגרמות ומפות חום של התכונות על גבי המפות. זוהי דרך ויזואלית, אינטואיטיבית ונוחה שמראה כיצד ה-data מפוזר על פני האזור שעליו אנו עובדים. ההיסטוגרמות ומפות החום מציגות את ריכוז נקודות העניין לפי סוג בכל אזור.

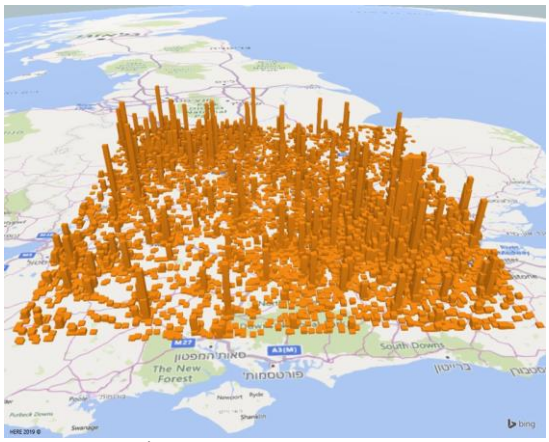
נציג בחלק זה דוגמאות בודדות, **ניתן לעיין יתר הגרפים מצורפים בנספח שבסוף המסמך**.

ברים (Bars):

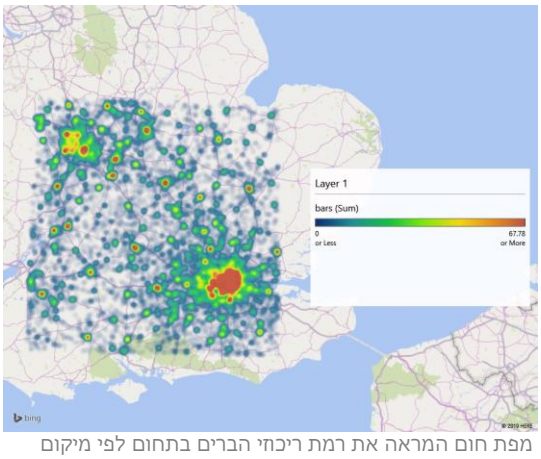
סוגי הברים שהוצאנו בעזרת ה-API של foursquare ופיזורם על המפה:



סוגי הברים והקוד שלהם ב-API של Foursquare - https://developer.foursquare.com/docs/resources/categories



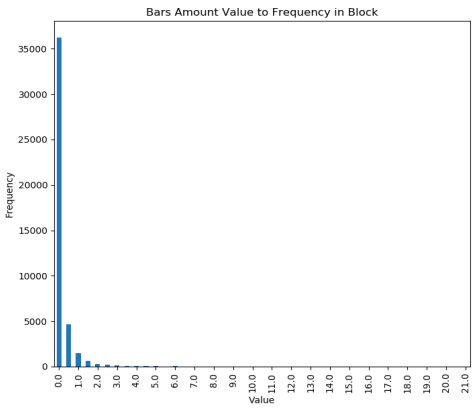
היסטוגרמה המראה את כמות הברים בתחום לפי מיקום

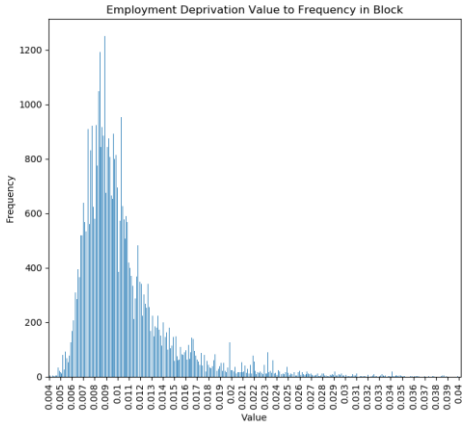


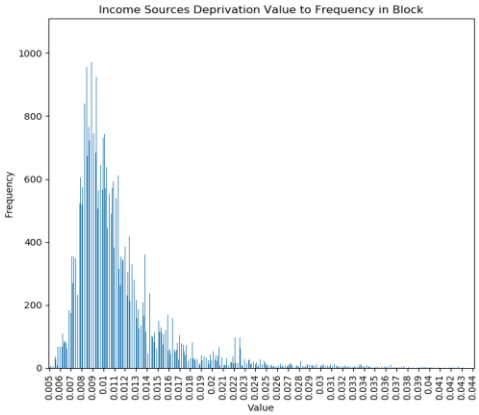
מפת חום המראה את רמת ריכוז הברים בתחום לפי מיקום

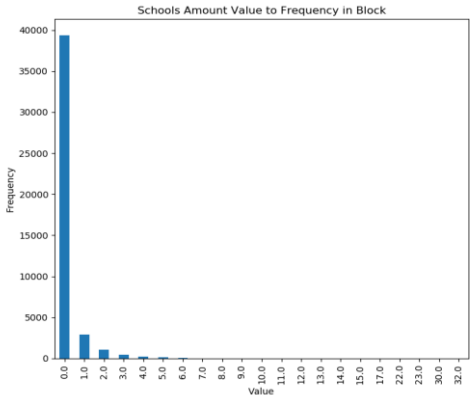
ניתן לראות שיש ריכוז מאוד גבוה של ברים בלונדון ובערים מרכזיות נוספות בהן ערכי ההיסטוגרמה על המפה גבוהים, ומידות החום במפת החום גבוהות.

בנוסף למפות החום הוצאנו היסטוגרמות המראות את ההתפלגות של כל אחת מהתכונות. עשינו זאת במטרה לבחור את שיטת ה-Scaling המתאימה ביותר עבור כל תכונה כפי שנסביר בפרק על Scaling.









ההיסטוגרמות הנ"ל הינן דוגמאות למספר הערכים המצויים בסט הנתונים לכל תכונה, עבור תכונות מספר בתאי הספר, מספר הברים וערכי המחסור עבור גישה להכנסה ותעסוקה (כאשר יתר ההיסטוגרמות מצורפות בנספח שבתחתית המסמך). מן הגרפים הנ"ל ניתן ללמוד על התפלגות של תכונה בסט הנתונים, כמפורט בפרק Scaling.

עיבוד מקדים של המידע

בחלק זה של הפרויקט נתאר את העיבוד המקדים של ה-data שביצענו. נפצל את חלק זה ל-2 חלקים, כאשר בחלק הראשון נסביר על Imputing (השלמת ערכים חסרים) ובחלק השני נסביר על Scaling (שינוי טווח ערכי התכונות).

Imputing

במאגר הנתונים הגולמי של ממשלת בריטניה שבו השתמשנו (מפורט בפרק של מקורות המידע) היו נתונים מיקומי תאונות הדרכים בכל שורה ושורה. בבניית מאגר הנתונים שלנו, המורכב ממשבצות, היו משבצות שבהן לא היו תאונות מכיוון שבמאגר הגולמי לא היו תאונות אשר מיקומן נפל בתוך משבצות אלה. כתוצאה מכך לא היה לנו במאגר שבנינו את ה-LSOA Code עבור אותן המשבצות (כי זהו נתון שהיה במאגר המקורי עבור התאונות הנתונות). עבור משבצות אלה ביצענו Imputing לLSOA Code על ידי מציאת הLSOA המוצמד לPostal Code הקרוב ביותר למרכז המשבצת ע"י Postcodes Api. בנוסף, בעזרת תכונת ה-LSOA code ביצענו imputing לכל התכונות הנגזרות ממנה עבור משבצות שבהן לא היו תאונות.

Scaling

ישנם אלגוריתמי למידה אשר בהם לגודל (Magnitude) ערכי התכונות יש השפעה על מידת ביצועם. דוגמא נפוצה היא KNN (המפורט בהמשך), עבורו נמדדת מטריקת מרחק בין דוגמאות שונות, אשר עבורה שוני גדול בין ערכי תכונות יכול להשאיף לאפס את מידת השפעת תכונות שפיזורן מצוי על פני Magnitude קטן בהשוואה. על מנת לפתור את הבעיה הנ"ל, נהוג לבצע Scaling לתכונות ככה שיפוזרו בין ערכים קבועים (למשל בין 0 ל1), או לפזרם גאוסיינית סביב מרכז זהה.

ישנן שיטות שונות לביצוע Scaling, אך נפוץ כי עבור תכונות המתפלגות בהתפלגות שקרובה להתפלגות יוניפרומית, משתמשים בScaling מסוג MinMax (על מנת לתחום אותם בין ערכים קבועים, [0,1] או [-1,1]), ע"י הנוסחא- $x_{scaled} = \frac{x - min(x)}{max(x) - min(x)}$ כאשר $min(x)$ הינו המופע הנמוך ביותר של תכונה x, ובהתאמה $max(x)$ הינו המקסימלי (בסט האימון).

לתכונות המתפלגות גאוסיינית נהוג לבצע Scaling מסוג Standardization, המכונה גם Z-score Normalization, המחושב ע"י הנוסחא- $x_{scaled} = \frac{x - \mu}{\sigma}$ כאשר μ הוא ממוצע ערכי התכונה וσ היא סטיית התקן (בסט האימון).

על פי הגרפים של פיזור התכונות, בחרנו לבצע Scaling מסוג Standardization לתכונות, שכן עבור כל התכונות, צורת ההתפלגות דמתה לגאוסיינית מאשר לנורמלית- ולכן ביצוע Standardization יפגע פחות ביחסים שבין התכונות.

הנדסת פיצ'רים

נדרשנו לבצע הנדסת פיצ'רים על המידע שהוצאנו מ-Foursquare (בתי ספר, פאבים ומקומות עניין תרבותיים). לאחר שליחת ה-request לקבלת המידע מ-Foursquare Places API קיבלנו ב-response אובייקט בפורמט json המכיל פירוט על המקומות הרלוונטיים במשבצת המבוקשת. על מנת לקבל תכונה של מספר המקומות במשבצת ביצענו סכימה של האלמנטים המוכלים ב-json.

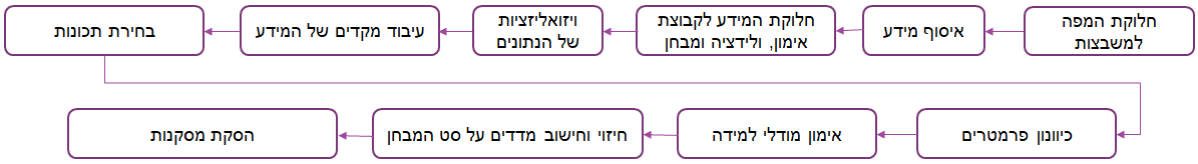
בנוסף, על מנת לקבל את הנתונים על מספר תאונות הדרכים במשבצת, ממוצע חומרת התאונות במשבצת, מספר הנפגעים במשבצת ומספר כלי הרכב המעורבים בתאונות במשבצת, היה עלינו לבצע עיבוד של המידע הגולמי המופיע בטבלת "מאגר תאונות הדרכים בבריטניה" (המתואר תחת מקורות המידע).

בטבלת המידע הגולמי, כל שורה ייצגה תאונת דרכים שנרשמה ע"י הרשויות, בצירוף הקואורדינטות שלה וה-LSOA Code של מיקום התאונה. לפיכך, היה עלינו לבצע סריקה על הטבלה הנ"ל ולשייך כל תאונה שהתרחשה במסגרת המרחב הנבדק למשבצת מתאימה על פי הקוארדינטות, וכך לחשב את התכונות הרלוונטיות לכל משבצת מהתאונות המשויכות לה.

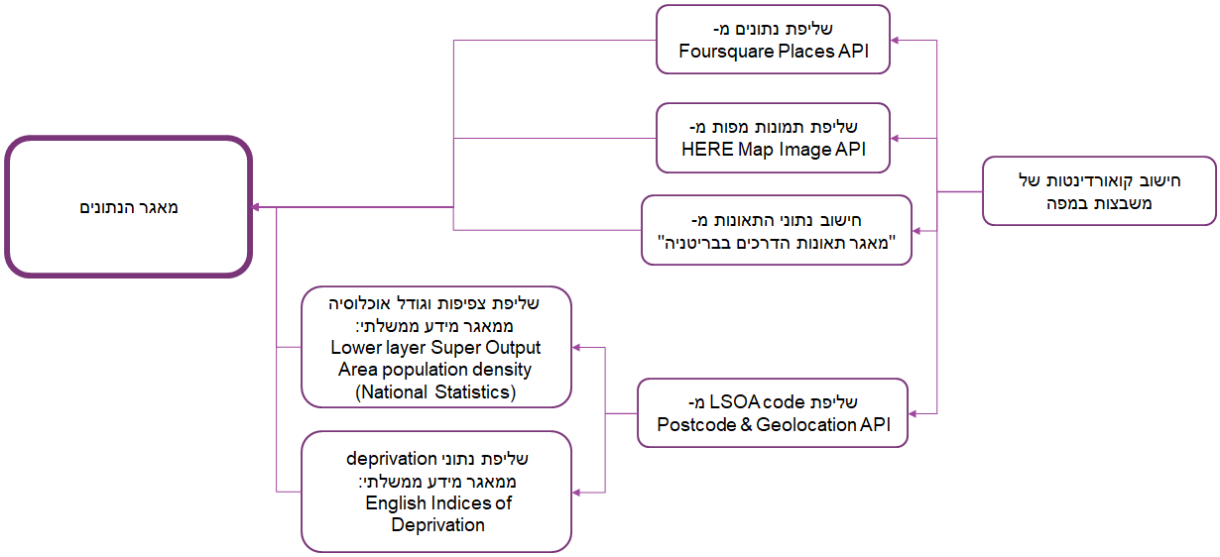
תיאור המערכת

בחלק זה נתאר את מבנה המערכת באמצעות דיאגרמות ונפרט בקצרה על רכיביה השונים.

דיאגרמה המערכת



דיאגרמת איסוף הנתונים



מבנה המערכת

חלוקת המערכת לרכיבים לוגיים:

חלוקת המפה למשבצות

קלט	קואורדינטות של הנקודה השמאלית התחתונה, הימנית העליונה, אורך ורוחב משבצת
פלט	קובץ csv בשם map_grid.csv המכיל קואורדינטות של המשבצות בתחום, המתוייגות ב-id מספרי
שם הקובץ	MapSplitter.py

איסוף הנתונים

שליפת נתונים מ- Foursquare Places API

קלט	קואורדינטות של פינה ימנית עליונה, שמאלית תחתונה וקטגוריה
פלט	מספר נקודות העניין מסוג הקטגוריה בתוך התחום
שם הקובץ	foursquareDataHandler.py

שליפת תמונות מפות מ- HERE Map Image API

קלט	קואורדינטות של פינה ימנית עליונה ושמאלית תחתונה
פלט	תמונת מפה של המשבצת
שם הקובץ	mapImagesHandler.py

חישוב נתוני התאונות מ-"מאגר תאונות הדרכים בבריטניה"

קלט	קובץ csv המכיל את המשבצות במפה
פלט	קובץ csv המכיל בכל שורה את נתוני התאונות עבור משבצת
שם הקובץ	DataGatherer.py

שליפת LSOA code מ- Postcode & Geolocation API

קלט	קואורדינטות במפה
פלט	רשימת קודי LSOA הקרובות לקואורדינטות הקלט
שם הקובץ	lsoaDataHandler.py

שליפת צפיפות וגודל אוכלוסיה ממאגר מידע ממשלתי: LSOA Area population density

קלט	קוד LSOA
פלט	שלשת שם איזור הLSOA, כמות האוכלוסייה וצפיפות האוכלוסייה
שם הקובץ	lsoaDataHandler.py

שליפת נתוני deprivation ממאגר מידע ממשלתי: English Indices of Deprivation

קלט	קוד LSOA
פלט	ערך הDeprivation המבוקש עבור קוד הLSOA
שם הקובץ	lsoaDataHandler.py

חלוקת המידע לקבוצת אימון, ולידציה ומבחן

קלט	dataframe עם נתוני התאונות - accidents_1km
פלט	חלוקה ל-3 סטים, אימון ולידציה ומבחן
שם הקובץ	DataPreparator.py

עיבוד מקדים של המידע

קלט	dataframes עם הנתונים - accidents_1km
פלט	dataframe עם הנתונים מעובדים train_processed.csv, validation_processed.csv, test_processed.csv
שם הקובץ	DataPreparator.py

בחירת תכונות

קלט	סט האימון, מודל רגרסיה, score function
פלט	הפיצ'רים שהאלגוריתם בחר כפלט מודפס
שם הקובץ	FeatureSelector.py

כיוונון פרמטרים

קלט	סט האימון, מודל רגרסיה, מילון המכיל את הפרמטרים הנבחרים.
פלט	ערכי הפרמטרים הנבחרים עבור המודל בקובץ params_results.txt
שם הקובץ	hyperparameterTuning.py

אימון מודלי למידה

קלט	מודל הרגרסיה, סט אימון, פרמטרי המודל
פלט	מודל מאומן (עבור רשת הנוירונים - נשמר כcheckpoint בשם final_model_r_{Experiment_Name}.pt)
שם הקובץ	MachineLearningEvaluation.py, DeepLearningTraining.py

חיזוי וחישוב מדדים על סט המבחן

קלט	מודל מאומן, סט המבחן
פלט	חיזוי מספר התאונות עבור הדוגמאות בסט המבחן וחישוב המדדים והדפסתם.
שם הקובץ	MachineLearningEvaluation.py, DeepLearningMain.py

תלויות (ספריות)

scikit-learn - מודלים, פונקציות score לחיזוי, פונקציות לצורך בחירת תכונות.

GeoPy - ספריה לעבודה עם נתונים גאוגרפיים.

Matplotlib - הצגת תוצאות בגרפים.

NumPy - אריתמטיקה של מטריצות ווקטורים.

PyTorch - רשתות נוירונים, למידה עמוקה.

את רשימת התלויות המלאה ניתן למצוא בקובץ **requirements.txt**

תיאור הניסויים

מדדי ההערכה בהם השתמשנו

על מנת לאמוד את טיב הפרדיקציות של המודלים השונים השתמשנו במדדי הערכה הבאים:

MSE – Mean Squared Error

מדד המתאר את ממוצע ריבועי הטעויות.

$$MSE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2$$

כאשר $\hat{y}_i$ זהו הערך החזוי של הדוגמא ה-i, ו- y_i זהו הערך האמיתי.

Mean Absolute Error

מדד המתאר את ממוצע הערכים המוחלטים של הטעויות.

$$MSE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i|$$

כאשר $\hat{y}_i$ זהו הערך החזוי של הדוגמא ה-i, ו- y_i זהו הערך האמיתי.

Explained Variance Score

מתאר את הפרופורציה לפיה מודל מתמטי מתאר את פיזור המידע בסט נתונים, כאשר הערך הטוב ביותר הוא 1, וערכים נמוכים יותר הינם טובים פחות. המדד מחושב ע"פ הנוסחא הבא:

$$explained\_variance(y,\hat{y}) = 1 - \frac{Var(y-\hat{y})}{Var(y)}$$

כאשר Var הוא השונות (ריבוע סטיית התקן) של הנתונים.

R2 Score

מתאר את הפרופורציה של שונות תוצאות החיזוי אשר מוסברות על ידי המשתנים הבלתי תלויים של המודל. הציון יכול להוות מדד לסבירות לחיזוי מוצלח של המודל על דוגמאות שלא נצפו. הציון הטוב ביותר הוא 1, כאשר תוצאות נמוכות יותר הן טובות פחות. ציון 0 יינתן למודל אשר יבצע חיזוי זהה לכל דוגמה המתקבלת כקלט לחיזוי.

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

כאשר $\hat{y}_i$ זהו הערך החזוי של הדוגמא ה-i, ו- y_i זהו הערך האמיתי.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

1. ניסויים לבחירת תכונות

על מנת לבדוק אילו תכונות הינן קריטיות לחיזוי, ובניסיון לאתר תכונות שניתן להסיר- ערכנו מספר ניסויים באמצעות שיטות שונות של Feature Selection.

SFS, SBS ו-BDS הינן שלוש וריאציות של Wrapper Methods, אשר פועלים בשיטת Sequential Feature Selection.

Wrapper Methods הינן שיטות Feature Selection אשר פועלות על ידי "עטיפת" מודל חיזוי באלגוריתם אשר מדרג את ביצועיו על גבי תתי קבוצות שונות של תכונות. ב Sequential Feature Selection, הכוונה שהתכונות השונות נבדקות סדרתית אחת אחר השנייה, בתוספת לתת קבוצה מוסכמת של תכונות.

Sequential Forward Selection – SFS

האלגוריתם מתחיל עם קבוצה ריקה של תכונות נבחרות, ובזו אחר זו מוסיף לתת קבוצת התכונות הנבחרות את התכונה שהוספתה לקבוצת התכונות הנבחרות הובילה לשגיאה הנמוכה ביותר על המודל הנבדק על סט הוולידציה. האלגוריתם עוצר כאשר כל התכונות צורפו לקבוצת התכונות הנבחרות, או כאשר כל הוספה של תכונה תוביל לעלייה בשגיאה של המסווג הנבחר על סט הוולידציה.

Sequential Backward Selection - SBS

האלגוריתם מתחיל עם קבוצה מלאה של תכונות נבחרות, ובזו אחר זו מחסיר מתת קבוצת התכונות נבחרות את התכונה שהחסרתה מקבוצת התכונות הנבחרות הובילה לשגיאה הנמוכה ביותר על המודל הנבדק על סט הוולידציה. האלגוריתם עוצר כאשר כל התכונות הוסרו, או כאשר כל החסרה של תכונה תוביל לעלייה בשגיאה של המסווג הנבחר על סט הוולידציה.

Bi-Directional Search - BDS

אלגוריתם המבצע הרצה מתחלפת של צעדי SFS ו-SBS אחד לעבר השני, עד להגעה להתכנסות לסט תכונות זהה. זה נאכף על ידי כך שתכונות שנבחרו על ידי צעד SFS לא יוחסרו על ידי צעד SBS, ובאופן דומה תכונות שהוחסרו על ידי צעד SBS לא יוספו על ידי צעד SFS.

הטבלאות להלן מייצגות את התכונות שנבחרו ע"י האלגוריתמים הנ"ל עבור הרגרסורים, בהתאם לשורה המתאימה בטבלה-

SVR (Support Vector Regressor), LR (Linear Regression), GB (Gradient Boosting Regressor), KNN (K-Neighbors Regressor), BR (Bayesian Ridge), RF (Random Forest Regressor)

כאשר העמודות מייצגות את התכונות, וסימון בירוק מסמל תכונה נבחרת.

SFS:

	income	bars	pop	density	health	living	school	crime	arts	edu	employ
SVR	X	X		X	X				X	X	
LR	X		X	X		X	X			X	X
GB	X	X	X	X	X	X	X	X		X	
KNN		X		X	X	X		X			
BR	X		X	X		X	X	X	X	X	X
RF	X	X		X	X	X	X	X			

SBS:

	income	bars	pop	density	health	living	school	crime	arts	edu	employ
SVR	X	X		X	X				X	X	
LR	X		X	X		X	X	X	X	X	X
GB	X	X	X	X	X	X	X	X			X
KNN	X			X		X	X	X		X	X
BR	X		X	X		X	X	X	X	X	X
RF	X	X	X	X	X	X	X	X	X		X

BDS:

	income	bars	pop	density	health	living	school	crime	arts	edu	employ
SVR	X				X		X	X	X		X
LR	X				X		X	X	X		X
GB	X				X		X	X	X		X
KNN	X				X		X	X	X		X
BR	X				X		X	X	X		X
RF	X	X				X	X	X		X	

נבחין כי עבור 3 השיטות מתקבלת הסכמה כמעט גורפת על חשיבות התכונות Income Deprivation, Crime Domination אולם ישנן תכונות כגון Population Density אשר נבחר עבור כל המודלים בשיטות SFS וSBS, אולם לא נבחר עבור אף מודל בBDS. כמו כן נשים לב כי BDS וSBS הציגו הסכמה יחסית על Employment Deprivation, בעוד SFS לא.
מלבד אישוש חשיבות Income Deprivation, Crime Domination לא ניתן לקבוע כי אף אחת מהתכונות הנ"ל אינה שימושית לחיזוי.

סיבה אפשרית לחוסר אחידות התכונות הנבחרות על ידי השיטות השונות היא כתוצאה מתכונות שונות שיכולת ההסקה מהן על תכונת המטרה (מספר התאונות) דומה עד שווה, ולכן בהינתן שנבחרה תכונה אחת, צירופה של תכונה אחרת לא תשפר את השגיאה ועלולה אף להגדילה.

על מנת לבדוק את האפשרות הנ"ל, ביצענו ניסוי נוסף בו השוואנו, באמצעות מדד Normalized Mutual Information[12], את מידת המידע המשותף בצירוף של כל צמד תכונות.
נבחין כי המדד נע בין 0 ל1, כאשר 1 מייצג יכולת מוחלטת להסקת תכונת אחת מהשנייה (ומכך, המידע המשותף בין כל תכונה לעצמה הוא 1). הטבלה להלן מייצגת את התוצאות של הרצת הניסוי.

	school	bars	arts_and_entertainment	population	population_density	crime_dom	education_dep	employment_dep	health_dep	income_dep	living_environment_dep
school	1.00	0.19	0.18	0.09	0.20	0.13	0.19	0.05	0.13	0.06	0.20
bars	0.19	1.00	0.21	0.10	0.19	0.13	0.19	0.05	0.14	0.06	0.20
arts_and_entertainment	0.18	0.21	1.00	0.10	0.18	0.13	0.19	0.05	0.13	0.06	0.19
population	0.09	0.10	0.10	1.00	0.78	0.87	0.92	0.64	0.88	0.66	0.93
population_density	0.20	0.19	0.18	0.78	1.00	0.81	0.87	0.55	0.82	0.57	0.87
crime_dom	0.13	0.13	0.13	0.87	0.81	1.00	0.95	0.68	0.90	0.70	0.95
education_dep	0.19	0.19	0.19	0.92	0.87	0.95	1.00	0.76	0.95	0.77	0.99
employment_dep	0.05	0.05	0.05	0.64	0.55	0.68	0.76	1.00	0.68	0.42	0.76
health_dep	0.13	0.14	0.13	0.88	0.82	0.90	0.95	0.68	1.00	0.71	0.95
income_dep	0.06	0.06	0.06	0.66	0.57	0.70	0.77	0.42	0.71	1.00	0.78
living_environment_dep	0.20	0.20	0.19	0.93	0.87	0.95	0.99	0.76	0.95	0.78	1.00

מניתוח הטבלה, ניתן לראות כי ישנן תכונות אשר מכילות מידע משותף אחת על השנייה, בייחוד בין תכונות הDeprivation השונות, יחד עם כמות וריכוז האוכלוסייה.

התכונות אשר מכילות את המידע המשותף הרב ביותר לפי המדד הנ"ל הן Living Environment Deprivation וEducation Deprivation, אשר קיבלו ציון של 0.99 מתוך 1.

[12] "sklearn.metrics.normalized_mutual_info_score — scikit-learn ...." http://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html. Accessed 21 Oct. 2019.

ואולם, מן התוצאות של הWrapper Methods שתוארו מעלה, נשים לב כי מלבד שיטת BDS בה לא נבחרו עבור מרבית ההרצות שתי התכונות הנ"ל, הן נבחרו במשותף במקרים רבים עבור אותם מודלים בשיטות SFS וSBS. משמעות הדבר שלאחר הוספת תכונה אחת לאחר השנייה (שכן האלגוריתם מצרף/ מחסיר תכונות סדרתית), התקבלה ירידה בשגיאה המתקבלת על סט הוולידציה.

מכך אנו מסיקים כי אף על פי שלפי מדד המידע המשותף, התכונות הנ"ל מכילות מידע משותף רב- יתכן כי במקרים מסוימים ניתן להשתמש בצירופם על מנת להגיע לתוצאת רגרסיה מדויקת יותר.

מהסיבה הנ"ל בחרנו לא להחסיר תכונות מהסט שנאסף, שכן כל התכונות שאספנו עוזרות בחלקן לכמות לא זניחה של רגרסורים בשיפור תוצאת החיזוי- ועל כן בהחסרתן אנו נפגע בתוצאה המתקבלת עבורם.

2. ניסויים לחיזוי מספר תאונות דרכים באמצעות אלגוריתמי למידה

על מנת לענות על הבעיה של חיזוי רמת הסיכון במשבצת גאוגרפית כפי שהעלנו בתחילת הדו"ח החלטנו להשתמש במספר אלגוריתמי רגרסיה כפי שנפרט בניסויים הבאים.

ראשית, ביצענו כיוונון פרמטרים עבור האלגוריתמים שהשתמשנו בהם. על מנת לבצע את הכיוונון החלטנו להשתמש במנגנון GridSearchCV של sklearn. מנגנון זה מאפשר לבצע כוונון פרמטרים אוטומטי על ידי מעבר על מספר קומבינציות אפשריות של ערכים מבין האופציות שהועברו לו.

כל בחינה של פרמטרים מתבצעת על ידי אימון ובחינה של המודל עם ערכי הפרמטרים הנבדקים בשיטת K-fold cross validation, תוך הערכה של הדיוק המתקבל. בחרנו להשתמש ב-$k = 5$ בעיקר משיקולי סיבוכיות זמן ריצה.

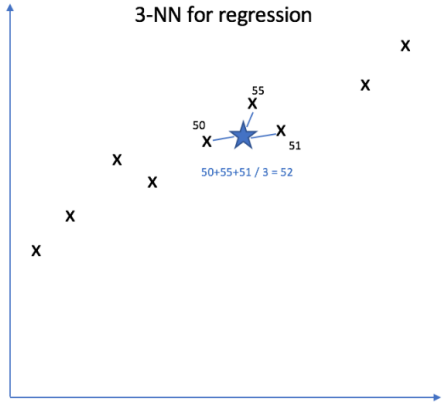
כך עבור כל אלגוריתם קיבלנו את הפרמטרים המיטביים שבחר האלגוריתם באופן אוטומטי.

חשוב לציין שמספר הקומבינציות האפשריות של הפרמטרים עלול להיות גדול מאוד, לכן בחרנו לתת לאלגוריתם של GridSearchCV רשימת פרמטרים מוגבלת אשר דרשה זמן חישוב סביר.

בנוסף לזאת חשוב לנו לציין שרוב האלגוריתמים שנציג כאן אלו אלגוריתמים שמרבית שימושינו בהם במהלך התואר והחלק התאורטי הנלמד עליהם היו למטרות קלסיפיקציה כמו למשל Random Forest, SVM, KNN. כחלק מהפרויקט למדנו שניתן להשתמש באלגוריתמים אלו למשימות רגרסיה ולכן נרחיב בקצרה עבור כל מודל כיצד ניתן לבצע באמצעותו רגרסיה.

מודל KNN Regression

מודל KNN לרגרסיה מחזיר בתור פרדיקציה את ממוצע ערכי k השכנים הקרובים ביותר. כמו שניתן לראות בגרף מצד שמאל, כאשר מסתכלים על 3 השכנים הקרובים, עושים ממוצע של ערכיהם וזה הערך החזוי.



התמונה לקוחה מ-
https://www.jeremyjordan.me/k-nearest-/neighbors

הפרמטרים שבחנו ב GridSearchCV הם:

- מטריקת מרחק: מרחק אוקלידי, מרחק מנהטן
- משקול דוגמאות: אחיד, לפי הופכי המרחק
- מספר שכנים: 1,2,3,4,5,6,10,20

הפרמטרים שנבחרו על ידי GridSearchCV הם:

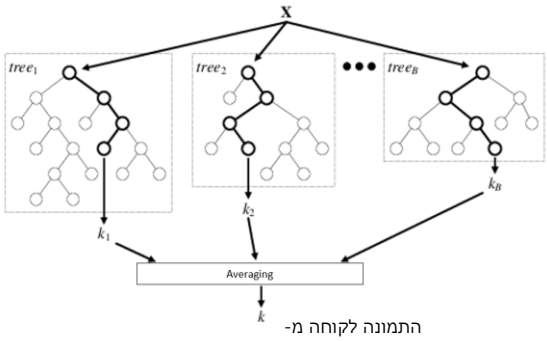
- מטריקת מרחק: מרחק אוקלידי

$$d(p,q) = d(q,p) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\ldots+(q_n-p_n)^2}$$

- משקול דוגמאות: בצורה אחידה. כלומר כל הדוגמאות ב"שכונה" ממושקלות באופן אחיד.
- מספר שכנים: 6. כלומר עושים ממוצע על 6 שכנים קרובים.

הבחירה של GridSearchCV ב-6 שכנים היא הגיונית מכיוון שאם מספר השכנים יהיה גדול מאוד זה גורם לכניסה של רעש הגורם לתוצאות פחות מדויקות. מצד שני, מספר קטן יותר של שכנים הוא פחות אינפורמטיבי כאשר מחשבים את ממוצע ערכי השכנים על מנת לקבל את תוצאת החיזוי.

מודל Random Forest Regression



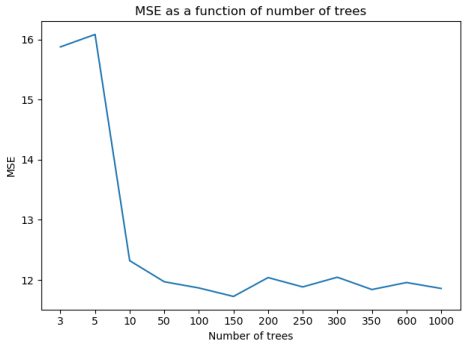
התמונה לקוחה מ-
https://www.researchgate.net/publication/301638643_Electromyographic_Patterns_during_Golf_Swing_Activation_Sequence_Profiling_and_Prediction_of_Shot_Effectiveness

במודל Random Forest לרגרסיה נבנים מספר של עצי החלטה מסט דוגמאות האימון ובזמן מבחן עבור דוגמה חדשה מתבצע ממוצע של תוצאות העצים השונים ביער.

במודל זה בחרנו לבדוק את השפעתם של מספר העצים ביער על החיזוי (על סט הוולידציה).

ניתן לראות בגרף מצד שמאל בציר ה-X את מספר העצים ביער ובציר Y את ה-MSE המתקבל עבור כל אחד מהם.

את ה-MSE הקטן ביותר (11.72) קיבלנו על ידי שימוש ב-150 עצים ביער.



ככלל, כיוון שהעצים פועלים כוועדה, ולכן מאפשרים צמצום שגיאות אחד של השני (אפילו אם למידתם חלשה), ככל שיש יותר עצים-המודל יהיה חזק יותר. ואולם עם העלייה במספר העצים, השיפור השולי קטן, כאשר כל עץ מוסיף לסיבוכיות האלגוריתם ולזמן הריצה.

הפרמטרים שבחנו ב GridSearchCV הם:

- עומק מקסימלי של עץ- 80, 90, 100
- מספר תכונות מקסימלי לפיצול- 2,3
- מספר דוגמאות מינימלי בעלה- 3,4,5
- מספר דוגמאות מינימלי לפיצול צומת פנימית- 8,10,12
- מספר מסווגים בוועדה - 100, 150, 200, 300

הפרמטרים שנבחרו על ידי GridSearchCV הם-

- עומק מקסימלי של עץ - 90
- מספר התכונות המקסימלי להפרדה בעץ - 3
- מספר מינימלי של דוגמאות בעלה - 8
- מספר דוגמאות מינימלי לפיצול צומת פנימית- 8
- מספר מסווגים בוועדה - 150

מודל Linear Regression

מודל מסוג Linear Regression מחפש את המישור הלינארי כך שסכום המרחקים הריבועיים בין הדוגמאות במרחב לבינו הוא הקטן ביותר (Least Squares). המודל פולט עבור וקטור x את תוצאת החיזוי על ידי חישוב הערך המתקבל על גבי המישור בנקודת ערכי תכונות הווקטור x.

מודל Support Vector Regression

רגסור SVR פועל באופן דומה למסווג SVM, על ידי מציאת Support Vectors (דוגמאות במרחב) אשר על ידי הגדרת על-מישור המוגדר על ידן מבצע הפרדה של מרחב הדוגמאות. במקרה הסיווג, האלגוריתם מאפשר הפרדה בעלת Margin מקסימלי, אשר העל-מישור שבמרכזן מגדיר את גבולות ההכרעה (במקרה של מסווג). ניתן להחליש את ההגבלה על ידי הגדרת גבולות "רכים" כך שמספר מסוים של דוגמאות אימון יוכלו לחרוג מהMargins המוגדרים על ידי הדוגמאות התומכות. ייחודו של הSVM במיפוי המידע למימד גבוה יותר, ובכך לאפשר למצוא הפרדה לינארית עבור בעיות שאינן ניתנות להפרדה לינארית במימד המקורי. פונקציית המיפוי מכונה Kernel, ומתקבלת כהיפרפרמטר של האלגוריתם.

במקרה של Support Vector Regression, האלגוריתם מחפש קו ישר במימד שהוגדר על ידי פונקצית הKernel, כך שכל דוגמא במרחב נמצאת במרחק שגיאה מוגבל ϵ ממנה. במקרה הנ"ל, הדוגמאות אשר הינן במרחק המקסימלי (ϵ) הינן ה Support Vectors, כיוון שמגדירות את על-המישור. לפיכך שימוש בSupport Vector Regression מאפשר שיערוך של פונקציית המטרה על ידי מציאת על-מישור במימד גבוה, אשר הינו בעל שגיאה של לכל היותר ϵ עבור דוגמא במרחב.

הפרמטרים שבחנו ב GridSearchCV הם:

- פרמטר קנס על שגיאה - 1,2,10
- גמא- פרמטר פונקציית הקרנל- הופכי מספר התכונות, $1e-7, 1e-4$
- ϵ= 0.1, 0.2, 0.4
- פונקציית קרנל - לינארי, פולינומיאלי, rbf - radial basis function

הפרמטרים שנבחרו על ידי GridSearchCV הם:

- פרמטר קנס על שגיאה - 10
- ϵ=0.1
- גמא- פרמטר פונקציית הקרנל- 0.0001
- פונקציית קרנל - rbf
- מספר איטרציות מקסימלי עד עצירה- 1000

מודל Bayesian Ridge Regression

מודל Bayesian Ridge Regression משערך MAP - Maximum a posteriori estimation תחת התפלגות גאוסיינית (תוך ההנחה שתגית המטרה y מפולגת גאוסיינית סביב Xw), כאשר הוקטור w ומשתנה מקרי α הם פרמטרים נלמדים, כלומר- $p(y|X,w,\alpha) = N(y|Xw,\alpha)$.

ההסתברות הפריורית לוקטור w נקבעת לפי -

$$p(w|\lambda) = N(w|0,\lambda^{-1}I_p)$$

כאשר הפרמטרים λ ו α נלמדים במשותף בעת אימון המודל.

הפרמטרים שבחנו ב GridSearchCV הם:

- פרמטרי התפלגות λ_1: $1e-6, 1e-5, 1e-7$
- פרמטרי התפלגות λ_2: $1e-6, 1e-5, 1e-7$
- פרמטרי התפלגות α_1: $1e-6, 1e-5, 1e-7$
- פרמטרי התפלגות α_2: $1e-6, 1e-5, 1e-7$
- טולרנטיות לשגיאה- $1e-3, 1e-2, 1e-4$

הפרמטרים שנבחרו על ידי GridSearchCV הם:

- פרמטרי התפלגות λ: $1e-05, 1e-07$
- פרמטרי התפלגות α: $1e-07, 1e-05$
- טולרנטיות לשגיאה - 0.01

מודל Gradient Boosting Regression

מודל Gradient Boosting Regression הינו מודל ensemble של weak learners (כלומר, מסווגים שדיוקם אינו גבוה משמעותית מניחוש). בהינתן מודל בשלב i בלמידה F_i, נגדיר את מודל השלב הבא $F_{i+1}(x) = F_i(x) + r(x)$. כיוון שהשאיפה היא שיתקיים $r(x) = y - F_i(x)$, המודל יתאים את r שיהווה קירוב של הפונקציה $y - F_i(x)$. לפיכך רעיונית, כל שלב באימון Gradient Boosting Regression מבצע תיקון של השגיאה המתקבלת מהשערוך שהתבצע בשלב האימון הקודם.

הפרמטרים שבחנו ב GridSearchCV הם:

- פרמטר קצב למידה-0.01, 0.02, 0.03
- עומק מקסימלי של העצים (מודלי הלמידה החלשים)- 4, 6, 8
- מספר העצים -100, 500, 600
- חלק הדוגמאות שילמדו כל מודל חלש - 0.9, 0.5, 0.2

הפרמטרים שנבחרו על ידי GridSearchCV הם:

- פרמטר קצב למידה- 0.03
- עומק מקסימלי של העצים (מודלי הלמידה החלשים)- 8
- מספר העצים - 600
- חלק הדוגמאות שילמדו כל מודל חלש- 0.9 (כלומר, כל מודל ילמד מ90% מדוגמאות האימון).

השוואת מדדים בין מודלי הלמידה על סט המבחן

	MSE	Mean Absolute Error	Explained Variance Score	R2 score
Linear Regression	90.06	7.37	1.71-	2.52-
Random Forest Regression	12.54	1.82	0.58	0.5
Support Vector Regression	38.21	4.86	0.14-	0.5-
Bayesian Ridge Regression	89.94	7.37	1.71-	2.54-
Gradient Boosting Regression	10.13	1.8	0.66	0.60
KNN Regression	8.06	0.95	0.68	0.68

ניתן לראות שתוצאות מודלי KNN, Random Forest ו-Gradient Boosting הן טובות יותר בניסוי זה בפער ניכר מיתר המודלים. נשים לב שהצלחתם של מודלים אלה לא תלויה בקיום קשר לינארי בין התכונות למשתנה המטרה.

מהצלחת אלגוריתם KNN אנו מסיקים כי מרחק אוקלידי קטן בין וקטורי דוגמאות מעיד על קרבה בין מספר התאונות של אותן הדוגמאות.

תוצאת המודל Linear Regression היא נמוכה מאוד (90.6 MSE) ואנחנו מסיקים מכך שאין קשר לינארי בין התכונות למספר התאונות, אם היה כזה סביר להניח שתוצאות מודל זה היו טובות יותר.

בהסתכלות על יתר המדדים- Mean Absolute Error, Explained Variance Score, R2 Score
ניתן להבחין שיחס הסדר בין תוצאות מדדים אלה למדד ה-MSE נשמר בהרצתם של האלגוריתמים השונים. כלומר אם נסתכל למשל על תוצאות מדדי אלגוריתם KNN, הן טובות יותר בכל המדדים מאשר שאר האלגוריתמים מהסיבות שציינו בפרק זה.

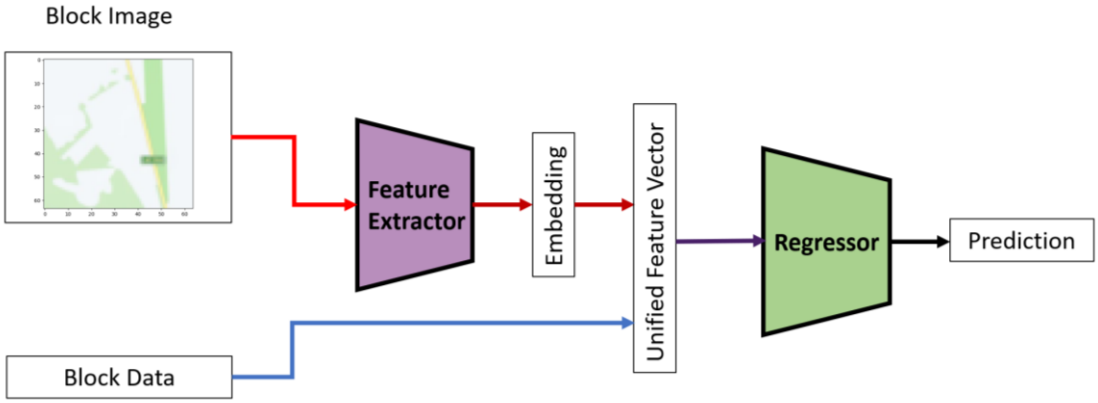
3. ניסויים לחיזוי מספר תאונת דרכים באמצעות למידה עמוקה

רשת נוירונים

על מנת להוסיף התחשבות בצורת מפת הכבישים, וכן בטופוגרפיה האיזורית ומבנה המפה הגאוגרפי, בחרנו לבדוק את השפעת הוספת תמונה של כל ריבוע נבדק במפה על ידי שליפת תמונה של המקום הנבדק תוך שימוש בפלטפורמה HERE, ושילובה לתוך אלגוריתם למידה. מודלי הלמידה על תמונה פורצי הדרך כיום משתמשים ברשתות נוירונים, אשר באמצעות שימוש בשכבות קונבולוציה[13] מונעות תלות במיקום המדויק של עצם בתמונה ובכך מונעים overfitting למבניות ספציפית.

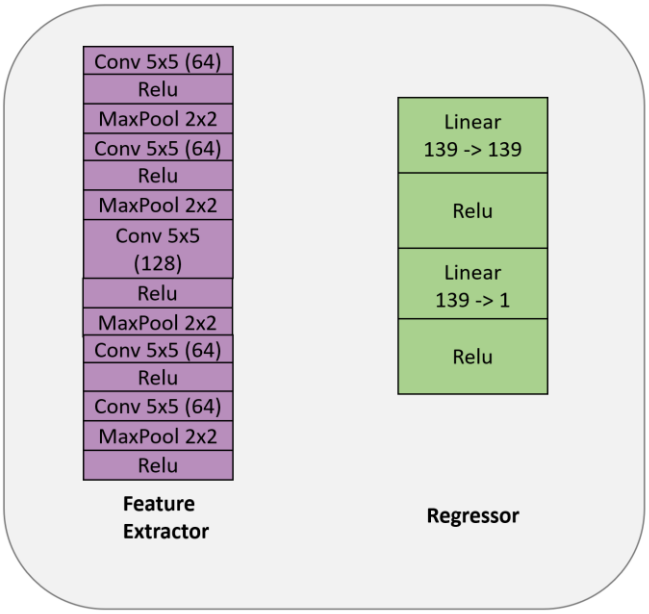
בנושא זה נציג ארכיטקטורה אותה פיתחנו על מנת לשלב את התכונות הנלמדות מתוך תמונה יחד עם תכונות מבוססות טבלה (שנאספו מיתר מקורות המידע עליהן התבססנו- ובהם אנו משתמשים ללמידה שאינה על רשת נוירונים), כך ששיערוך מספר התאונות יוכל להתחשב בכל התכונות שנאספו יחדיו.

מבנה זרימת הרשת



רשת הנוירונים מקבלת כקלט, במבנה של Mini- Batch, תמונה גולמית (מHERE, כמתואר מעלה), ושורה מתאימה למשבצת מטבלת התכונות. בטרם ההרצה התמונה עוברת טרנספורמציה לחיתוכה לריבוע ממורכז מדויק, וכן הורדת מימד ע"י צמצום הרזולוציה. (כאשר מידת הצמצום האופטימלית נבדקה בניסוי. בדיאגרמה- 64 על 64 פיקסלים). את התמונה אנחנו מעבירים דרך רשת נוירונים אותה אנו מכנים Feature Extractor שכן מטרתה היא צמצום התמונה ממימד המקור לכדי וקטור תכונות באורך שנקבע ל128. את וקטור התכונות, אותו אנו מכנים Embedding (שכן זהו Embedding של התמונה ל-Latent Dimension מוקטן), אנו מאחדים יחד עם וקטור התכונות המעובד מן הטבלה, לקבלת וקטור תכונות מאוחד. את הוקטור הנ"ל אנו מעבירים ברשת נוירונים המכונה Regressor, אשר מקבלת כקלט את וקטור התכונות המאוחד ופולטת מספר אי-שלילי המהווה שיערוך של מספר התאונות במשבצת הנ"ל.

[13] "Object Recognition with Gradient-Based Learning Yann ...." http://yann.lecun.com/exdb/publis/pdf/lecun-99.pdf.



הרכב רשת הנוירונים

את מבנה רשת הFeature Extractor ביססנו רעיונית על הארכיטקטורה המוכרת LeNet14, תוך העמקתה והתאמתה למבנה התמונות אותן אנו מספקים. רשת הRegressor מוגדרת בהתאם למספר התכונות הכולל המתקבל (11 שנאספו ידנית, ועוד 128 תכונות נרכשות מן התמונה). השימוש בRelu כשכבה אחרונה אוכף את אי השליליות של התוצאה (שכן מספר תאונות במשבצת הוא אי שלילי), יחד עם אי הגבלת הפלט באמצעות חסם עליון- כמו במקרה של שימוש בשכבת Sigmoid.

פרטי מימוש

על מנת לממש את הרשת הנ"ל השתמשנו בפלטפורמה PyTorch. הארכיטקטורה היחודית הצריכה כתיבת מחלקת Dataset חדשה אשר מבצעת את שליפת התמונה יחד עם שליפת השורה מבסיס הנתונים, ועיבודן לטנזורים בפורמט PIL המוכר על ידי PyTorch על מנת לאפשר את ההרצה. לבסיס הנתונים הנ"ל קראנו CombinationDataset, מאחר וכשמו הוא מאחד מאגר נתונים מסוג אוסף תמונות, ומאגר נתונים מסוג טבלת נתונים. המודל מתאמן במשך 100 אפוקים על סט האימון, והאיטרציה עבורה מתקבל ערך MSE Loss מזערי נבחר לשימור בcheckpoint ונרשמת תוצאתו עבור סט הולידציה.

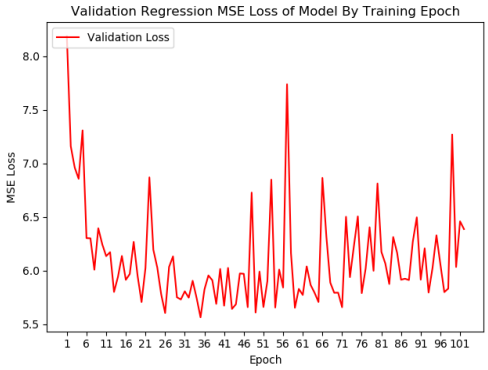
השתמשנו בAdam Optimizer לקידום הפרמטרים, עם היפר פרמטרים סטנדרטיים:
5-betas=(0.9, 0.999), lr=0.001, weight_decay=1e
כאשר weight_decay בתפקידו מונע את התנפחות ערכי הפרמטרים- ומכוון לשינוי היחס בניהם במקום הגדלה שלהם עד לגבולות הייצוג המספריים.

[14] "LeNet - Yann LeCun." http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf.

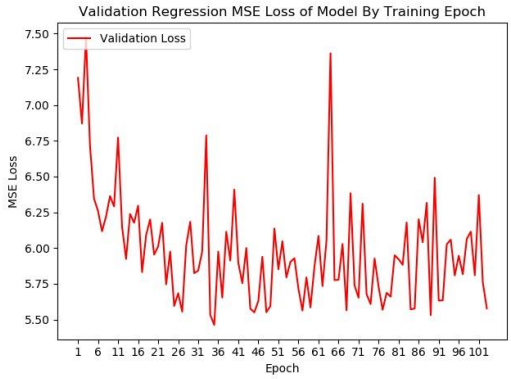
ניסויים

בגרפים הבאים מוצגת ההתקדמות של ערך הMSE Loss ע"פ הערכים החזויים על סט הוולידציה, לכל Epoch של אימון. כל גרף מתאר את תוצאות הבדיקה הנ"ל עבור תמונות ברזולוציה שונה, וכן בדיקת בקרה אשר אינה מבצעת למידה על תמונות.

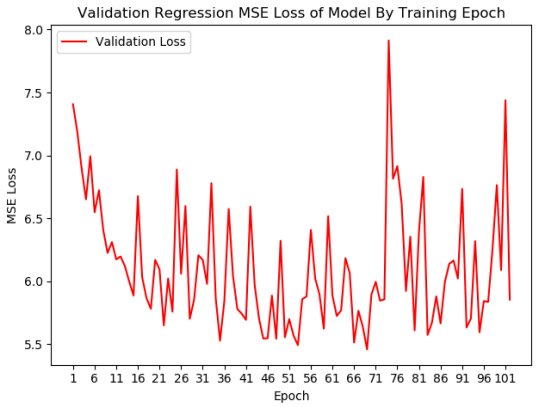
למידה עם תמונות בכיווץ לpixel*32pixel:32



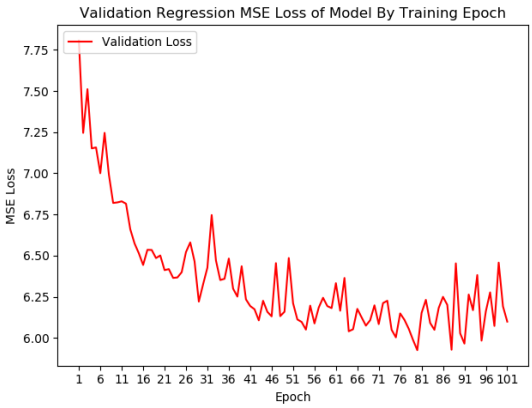
למידה עם תמונות בכיווץ לpixel*64pixel64:



למידה עם תמונות בכיווץ לpixel*128pixel:128



למידה ללא תמונות:



מדדי הרצת המודל על סט המבחן

	MSE	Mean Absolute Error	Explained Variance Score	R2 score
ללא שימוש בתמונות	7.62	0.93	0.70	0.70
שימוש בתמונות ממימד 32*32	7.18	0.88	0.71	0.71
שימוש בתמונות ממימד 64*64	7.03	0.84	0.72	0.72
שימוש בתמונות ממימד 128*128	6.56	0.80	0.74	0.74

נשים לב כי ניתן לראות שעבור כל המדדים, התוצאות הטובות ביותר מתקבלות עבור שימוש בתמונה ברזולוציה חדה יותר, כאשר ירידה באיכות התמונה ועד להגעה להעדר תמונה בסיווג- מפחיתה את הביצועים של המודל ומביאה איתה לירידה בביצועים, כפי שמדגימות המטריקות. מכאן אנו מסיקים שתצלום מפת האיזור מהווה כשלעצמה תכונה שאינה מיותרת בחיזוי מספר התאונות- ומאפשרת להביא את ביצועי רשת הנוירונים (המציגה בפני עצמה ביצועים טובים ביחס למודלים האחרים שנבדקו) לכדי שיאים חדשים.

4. ניסויים לבדיקת השפעת נתוני תאונות עבר על חיזוי מספר תאונות הדרכים

בניסוי זה רצינו לבדוק האם ניתן לשפר את רמת הדיוק של מספר התאונות הצפוי. השערתנו הסבירה היא שאיסוף ושימוש בנתוני השנה הקודמת על התאונות בכל משבצת תשפר באופן משמעותי את החיזוי של מספר התאונות. לצורך הבהרה- עד עכשיו השתמשנו בנתוני התאונות משנת 2018 וכעת אנו רוצים לבדוק כיצד שימוש בנתונים התאונות משנת 2017 יתרום לחיזוי.

הנתונים אותם אספנו מהשנה הקודמת על התאונות:

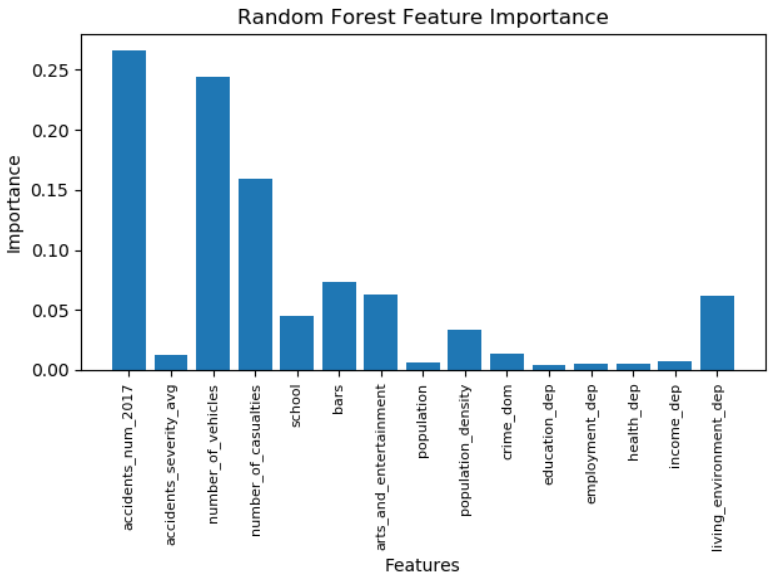
- מספר התאונות במשבצת
- ממוצע חומרת התאונות במשבצת
- מספר הנפגעים במשבצת
- מספר כלי רכב המעורבים בתאונות במשבצת

בניסוי זה ביצענו שוב הכנה מוקדמת של ה-data והרצנו פעם נוספת את GridSearchCV כמו בניסוי השני עבור כל אחד מהאלגוריתמים (מכיוון שסט הנתונים השתנה). לאחר מכן הרצנו שוב כל אחד מהאלגוריתמים ואלו התוצאות שהתקבלו על סט המבחן:

	MSE	Mean Absolute Error	Explained Variance Score	R2 score
Linear Regression	12.29	2.6	0.52	0.51
Random Forest Regression	3.1	0.91	0.88	087
Support Vector Regression	14.22	3.41	0.82	0.44
Bayesian Ridge Regression	12.26	2.6	0.52	0.51
Gradient Boosting Regression	3.36	1.18	0.89	0.86
KNN Regression	2.81	0.67	0.88	0.88
Neural Network (128pix*128pix)	2.59	0.66	0.89	0.89

ניתן לראות שיפור משמעותי במדדים. כלומר, אכן ניתן לאשש את ההשערה ולומר שנתוני העבר משפרים באופן משמעותי את חיזוי מספר התאונות. למשל, מקום שהיה בעבר מועד לתאונות דרכים, בסבירות די גבוהה יהיה גם בעתיד מקום המועד לתאונות דרכים (עד שלמשל יעשו פעולות משמעותיות על מנת למזער את מספר התאונות באותו האזור). כנ"ל לגבי מקום שבעבר לא היו בו תאונות או שהיו בו מספר מועט מאוד של תאונות.

הדיאגרמה הבאה ממחישה בצורה טובה את חשיבות הפיצ'רים שהוספנו לחיזוי מספר התאונות. הוצאנו את הפיצ'רים החשובים מאלגוריתם ה- Random Forest Regressor ואלו התוצאות שהתקבלו.



ניתן לראות שמספר התאונות במשבצת מהשנה הקודמת (2017_accidnets_num), מספר כלי הרכב אשר היו מעורבים בתאונה במשבצת בשנה הקודמת (number_of_vehicles) ומספר הנפגעים במשבצת בשנה הקודמת (number_of_casualties) קיבלו חשיבות גבוהה באופן משמעותי משאר הפיצ'רים.

עם זאת, חשוב לנו להדגיש שמטרה חשובה בפרויקט היא לתת הערכה כיצד שינוי אזורי (הוספת בתי ספר, הוספת צומת לכביש, הוספת גשר, הוספת פאבים וכדומה) ישפיעו על הסיכון לתאונות דרכים. לכן, ברגע שהוספנו את נתוני התאונות מהשנה שעברה אמנם הדיוק בחיזוי של מספר התאונות עלה, אך משקלי שאר הפרמטרים הקשורים לפיתוח האזורי ירדו באופן משמעותי, כלומר האלגוריתמים פחות מושפעים מקלטים שבהם המשתמש רוצה לבדוק כיצד שינוי בפיתוח אזורי ישפיע על מספר תאונות הדרכים. למעשה, על מנת שאלגוריתמי הלמידה יוכלו להשתמש בצורה אמיתית בנתוני תאונות הדרכים של העבר יחד עם הנתונים שנאספו מההווה (מספר מוקדי העניין, נתוני deprivation) יש צורך בצירוף נתוני deprivation ומוקדי העניין מהעבר.

כיוון שנתוני מוקדי העניין נאספו על ידנו השנה, ולא נגיש אלינו מנגנון המאפשר איסוף מידע שכזה משנים קודמות אשר לא כולל את נתוני ההווה (שאילתה על מקומות העניין מחזירה את אלו שקיימים ביום איסוף המידע), לא יכולנו לספק את הנתונים הללו במסגרת הפרויקט. ניתן לבצע הרצה חוזרת של איסוף הנתונים בשנה עתידית, ובכך להשתמש בנתוני העבר (נתוני ההווה מהפרויקט שלנו) על מנת לבצע למידה אפקטיבית אשר מתחשבת בתאונות העבר, וכן מבצעת הסקה על הקשר בין מספר מוקדי העניין, והשינוי בערכי deprivation, לבין מספר תאונות הדרכים בעתיד.

בנוסף לזאת, ניתן לראות שמשימת החיזוי של מספר תאונות הופכת להיות מדויקת יותר בשימוש בנתוני השנה שעברה, בניסוי זה רצינו להיווכח בכך ולהראות את התוצאות.

סיכום הפרויקט

מסקנות

בניסוי ברשתות הנוירונים הצגנו כי מתקבל שיפור בתוצאות החיזוי ברשת הנוירונים שנבחנה, בבדיקה בין שימוש במידע ללא התמונות אל מול שימוש במידע בצירוף התכונות- וכן התקבל שיפור בכל עליית רזולוציה של התכונות (עד למקסימלית שנבדקה). אנו מסיקים מכך כי תמונות מפה אכן מכילות מידע אשר הינו שימושי לצורך חיזוי מספר התאונות אשר יתרחשו במקום מסוים. מהשיפור עם העלייה ברזולוציית התמונה, אנו מבינים כי המודל עושה שימוש בפרטים קטנים הנמצאים בתמונות, ולומד לא רק צורות כלליות של מבנה הכבישים, אלא יתכן ומזהה סימוני כבישים מיוחדים או רחובות צרים במיוחד. המסקנה הזו משמחת, שכן אחת ממטרות הפרויקט הייתה לאפשר למתכנני ערים לבדוק השפעת וריאציות שונות של תכנון שכונות על מספר תאונות הדרכים במקום, ואם להסתמך על תוצאות הניסויים- נראה כי הדבר אפשרי.

בבחינת מודלי הלמידה השונים בניסוי 2, ראינו כי מרחב הדוגמאות אינו פשוט ללמידה, וכי אינו לינארי. מחצית ממודלי הלמידה שנבדקו בחלק זה הציגו קושי רב בלמידת פונקציית המטרה, והציגו תוצאות לא מוצלחות. בלטו מתוכם שלושה מודלי למידה שאינם מבוססים על לינאריות מרחב הדוגמאות, הרי הם KNN, GradientBoosting וRandomForest. מתוך שלושת המודלים הללו, רק KNN אינו מסווג הפועל בשיטת וועדה. הדבר מדגים את הכוח של וועדות בביטול החולשות המקומיות של מודלים בודדים אשר עשויים לעבור "התמקצעות" ברגרסייה של חלק מסוים ממרחב הדוגמאות, תוך הזנחת חלקים אחרים. מהצלחתו של KNN למדנו כי קיימת מידת רציפות מסוימת בנטייה למספר תאונות הדרכים במקום מסוים, ומסיבה זו דוגמאות הקרובות אוקלידית בתכונותיהן הן בעלות מספר דומה של תאונות דרכים זו לזו.

בבחינת מודלי הלמידה השונים על המידע שנאסף בצירוף נתוני התאונות משנה הקודמת לזו של מספר התאונות הנלמד (2018), ראינו כי קיים קשר הדוק בין מספר התאונות שהתרחשו במקום מסוים בשנה שעברה, לאלו שהתרחשו בו שנה לאחר מכן. הדבר אינו מפתיע, אך ראינו כי במקרה זה- ההתחשבות בשאר הנתונים שנאספו נעשתה מזערית. כפי שהסברנו בפרק הניסויים, אנו מבינים כי על מנת להתחשב בתכונות שאספנו משנת 2018, יחד עם מספר התאונות שנאסף ב2017, על בסיס הנתונים להכיל את גרסאות כל שאר התכונות מ2017- ובכך ללמוד את השינוי שתכונות אלו מביאות על מספר תאונות הדרכים. עקב מגבלה טכנולוגית ביכולת גרסאות אלו של התכונות שאספנו (כדוגמת הנתונים שנאספו מFoursquare על מספר מקומות העניין), לא יכולנו לספק את הנתונים הללו- אך אפשרות זו מראה פוטנציאל רב לצורך מחקר עתידי.

כאשר בדקנו את השפעת צירופן של התכונות השונות שאספנו על מידת השגיאה שהתקבלה בחיזוי, ראינו כי גם צירוף תכונות שבאופן רציונלי מידת השפעתם על תאונות דרכים אינה ברורה, כגון נגישות להכנסה ונוכחות פשיעה, הובילו לשיפור ביצועים עבור חלק גדול מן המסווגים להם הם צורפו. הדבר מעיד כי גורמים שונים שאינם טריוויאליים עשויים להשפיע על אופי האוכלוסייה במקום בצורות שישפיעו על מספר תאונות הדרכים, ולכן כאשר נבדקות שיטות להפחתת מספר

תאונות הדרכים במקום מסוים, יתכן וכדאי לשקול שיפור אלמנטים עקיפים בסביבה המקומית. אנו מקווים כי בעתיד שימוש במערכת המבוססת על עקרונות דומים לזו שפיתחנו תאפשר לגורמים בעלי השפעה למצוא באופן זה דרכים יצירתיות להפחתת מספר תאונות תוך שימוש בשיטות עקיפות.

קשיים

- כאשר התחלנו לעבוד על הפרויקט, Google Maps היה המקום הראשון והמוכר אליו פנינו על מנת להוציא מידע על סמך מיקום גאוגרפי. עבור הוצאת מידע על מקומות עניין לצורך בניית התכונות השתמשנו ב-Google Maps Places API. לאחר שחקרנו קצת יותר לעומק את השירות שהם מספקים גילינו שמספר השליפות מהשרתים מאוד מוגבל ומעבר לכמות זאת יש צורך לשלם סכום לא קטן של כסף עבור כמות השליפות הגדולה שנדרשה, לכן היה עלינו לחקור לא מעט על מקורות מידע אחרים שיספקו מידע מהימן.
- במהלך עבודתנו עם ה-APIs השונים היינו מוגבלים לתנאי השימוש של כל אחד מה-APIs. חלקם הגבילו את כמות הבקשות לשנייה, חלקם הגבילו את כמות הבקשות ביממה ובחלקם היו תנאים שונים על סוגי data שונים ששולפים מהם.
בפרויקט מסוג זה, הדורש כמויות גדולות של data זהו אתגר לא קטן למצוא מקורות מידע חינמיים המאפשרים כמות שליפות גדולה ותנאי שימוש נוחים. בסופו של דבר הצלחנו למצוא מספר מקורות מידע העומדים בדרישות, כפי שתיארנו בפרק מקורות המידע. עם זאת היה עלינו לעקוב ולנתר את השליפות בכל עת על מנת לעמוד בתנאים.
- בניסוי הרביעי כאשר בדקנו את השפעת נתוני תאונות עבר על החיזוי רצינו לספק למודל גם נתוני עבר עבור שאר התכונות (כמו בתי ספר, פאבים ועוד). מכיוון שה-APIs שעבדנו איתם סיפקו מידע עדכני לעת שליפת המידע לא הייתה לנו אפשרות לקבל מידע עבר עבור התכונות הנוספות. אנו סבורים שהזנת מודלי הלמידה בנתונים אלו הייתה משפרת את הביצועים מכיוון שמידע המצוי בשינויים בתכונות בין העבר להווה (כמו למשל שינוי במספר הפאבים) יכול להוסיף רובד נוסף של מידע למודלי הלמידה ובזאת להביא לשיפור ההסקה.
- רגרסיה לעומת קלסיפיקציה- פתרון בעיית חיזוי מספר התאונות לפי משבצת גאוגרפית כבעיית רגרסיה זהו אתגר לא פשוט. אנו משערים שסיווג בעיה זאת כבעיית קלסיפיקציה של תהיה תאונה/לא תהיה תאונה או חלוקת מספר התאונות למספר מוגבל של תחומים הייתה מאפשרת הצגת מדדים טובים יותר בדוח. עם זאת, מצאנו עניין ואתגר בפתרון בעיה זאת כבעיית רגרסיה ולמדנו מכך רבות.

רעיונות לשיפורים עתידיים

במהלך העבודה על הפרויקט העלנו מספר רעיונות לשיפור ופרויקטי המשך שיכולים להביא לשיפור בחיזוי.

- את המידע על נקודות העניין שהוצאנו מה-APIs השונים ניתן להוציא כל תקופת זמן מוגדרת ולהזין את נתונים אלו למודלים השונים על מנת שיהיה להם מידע מעודכן ורלוונטי שיכול לסייע בהסקה על מספר תאונות הדרכים הצפוי. במהלך הפרויקט ביצענו את שליפות הנתונים במשך מספר ימים רצופים ובאופן חד פעמי מכיוון שמשאבי מידע אלא הינם חינמיים ומגבילים את מספר הבקשות כפי שתארנו בפרק מקורות המידע.
- בחינה של תכונות נוספות שיכולות להביא לשיפור בחיזוי. ניתן להוציא מידע נוסף מה-APIs שהשתמשנו בהם (וגם מאחרים) על תכונות נוספות עבור כל משבצת. על ידי כך ניתן למצוא קשרים חדשים שיכולים להוסיף מידע על משתנה המטרה ולהביא לשיפור בתוצאות החיזוי.
- הרחבת ארכיטקטורת הלמידה העמוקה שבנינו עבור פתרון הבעיה. ניתן לחשוב על שיפור לארכיטקטורה הייעודית שבנינו על מנת לשפר את תוצאות החיזוי.

  ניתן למשל לחשוב על רשת Regressor מורכבת על ידי העמקתה.
- ניתן לבצע את החיזוי למספר תאונות הדרכים בעזרת אלגוריתמי קלסיפיקציה ולהשוות לתוצאות הרגרסיה. פתרון אפשרי הוא לבצע חלוקה של מספר התאונות החזוי לתחומים ולנסות לסווג כל משבצת לאינטרוול אחר בתחום.
- ניתן לבצע איסוף של תמונות לווין לפי המשבצות ולהזין כקלט לאלגוריתם הלמידה העמוקה שבנינו במקום תמונות המפה שבהן אנו השתמשנו. אנו מאמינים שבתמונות הלווין יש מידע נוסף כמו למשל מאפייני תוואי השטח אשר יכול להוסיף מידע לרשת.
- בדיקת המודל שבנינו במדינות נוספות. כלומר, ניתן לבדוק האם מודל שאומן במדינה מסוימת (במקרה שלנו בריטניה) יכול להסיק בצורה טובה גם על מדינות ואזורים נוספים בעולם ועל ידי כך לגלות גורמים חזקים שמשפיעים על כמות תאונות הדרכים בעולם.

מסקנות אישיות

- עבור פרק הניסוי על רשת נוירונים ולמידה עמוקה, פיתחנו עצמאית ארכיטקטורה ייעודית לבעיה שאינה "בעיית צעצוע", כאשר זה היה נסיוננו הראשון בעבודה מול בעיה שכזו.

  העבודה על רשת הנוירונים הייתה מאתגרת, כאשר בטרם הרצתה לא הייתה לנו הוכחה לכך שצירוף תמונות המפות לצורך החיזוי אכן יוביל לשיפור לו אנו מקווים בשגיאת החיזוי. לבסוף, תוצאות הניסוי תאמו בדיוק לציפיותינו. מלבד השיפור שבא עם צירוף התמונות, נראה אף שיפור בהעלאת רזולוציית התמונה המצורפת מנמוכה (32*32 פיקסלים), בינונית (64*64 פיקסלים) לגבוהה (128*128 פיקסלים). שמחנו לראות שעבודנו הרבה על רשת הנוירונים הייעודית השתלמה, ושהיא אכן שיפרה את תוצאות החיזוי באופן ניכר.

- במהלך הקורסים העוסקים בלמידה בטכניון, לא הורחב באופן משמעותי שימוש ברגרסיה באלגוריתמים שהשתמשנו בהם בניסוי השני העוסק בחיזוי מספר תאונות הדרכים באמצעות אלגוריתמי למידה. רוב המידע שרכשנו במהלך התואר בטכניון על אלגוריתמים אלה היה עבור פתרון בעיות קלסיפיקציה.
  לשם כך היה עלינו לחקור על הצורה בהם עובדים אלגוריתמים אלו לביצוע משימות רגרסיה. הגדלנו את ארגז הכלים שלנו באמצעות הרחבת הידע שלנו על אלגוריתמים אלו והאפשרויות שהם מספקים בפתרון בעיות מסוגים שונים.

- בפרויקט זה עסקנו בניסיון לפתור ולהתמודד עם בעיית למידה מציאותית באופן עצמאי. כחלק מהעבודה היה עלינו לתכנן את המערכת, לחפש מקורות מידע, לדלות את המידע, לממש את המערכת, לבנות תוכנית ניסויים ועוד. ניסיון זה תרם לנו רבות בהבנה תאורטית ופרקטית של הרכיבים השונים במערכת, בבניית ארכיטקטורה המתאימה לפתרון הבעיה ושימוש בארכיטקטורות ומודלים מוכרים לפתרון הבעיה. אנו סבורים שבעתיד כאשר ניגש לפתור בעיית למידה, סט הכלים שרכשנו תוך כדי העבודה בפרויקט יתרום לנו רבות.

נספחים

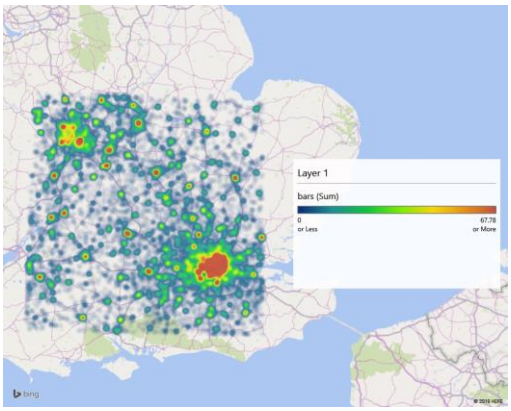
מפות חום והיסטוגרמות גאוגרפיות תלת מימדיות

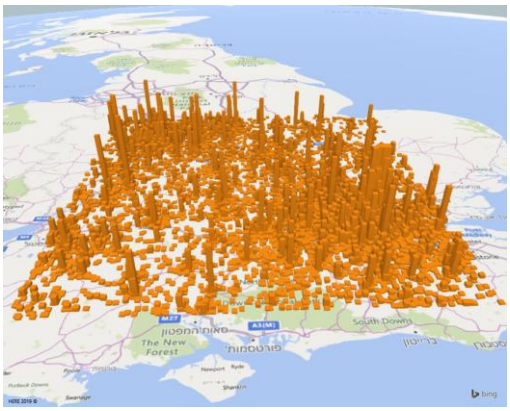
ברים (Bars)

סוגי הברים שהוצאנו בעזרת ה-API של foursquare ופיזורם על המפה:

Hotel Bar 4bf58dd8d48988d1d5941735	Speakeasy 4bf58dd8d48988d1d4941735	Wine Bar 4bf58dd8d48988d123941735	Champagne Bar 52e81612bcbc57f1066b7a0e Supported countries: GB
Karaoke Bar 4bf58dd8d48988d120941735	Sports Bar 4bf58dd8d48988d11d941735	Beach Bar 52e81612bcbc57f1066b7a0d	Cocktail Bar 4bf58dd8d48988d11e941735
Pub 4bf58dd8d48988d11b941735	Tiki Bar 56aa371be4b08b9a8d57354d	Beer Bar 56aa371ce4b08b9a8d57356c	Dive Bar 4bf58dd8d48988d118941735
Sake Bar 4bf58dd8d48988d11c941735	Whisky Bar 4bf58dd8d48988d122941735	Beer Garden 4bf58dd8d48988d117941735	

סוגי הברים והקוד שלהם ב-API של Foursquare - https://developer.foursquare.com/docs/resources/categories





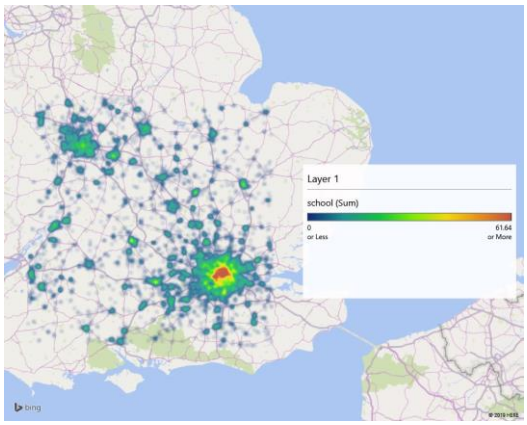
ניתן לראות שיש ריכוז מאוד גבוה של ברים בלונדון ובנוסף בערים נוספות בהם ערכי ההיסטוגרמה על המפה גבוהים, ומידות החום במפת החום גבוהות.

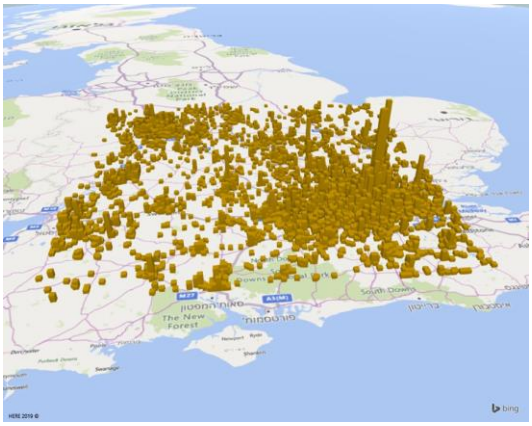
מקומות לימוד (Schools):

סוגי מקומות הלימוד שהוצאנו בעזרת ה-API של foursquare והפיזור על המפה:

Circus School 52e81612bcbc57f1066b7a43	Elementary School 4f4533804b9074f6e4fb0105	Language School 52e81612bcbc57f1066b7a48	Nursery School 4f4533814b9074f6e4fb0107
Cooking School 58daa1558bbb0b01f18ec200	Flight School 52e81612bcbc57f1066b7a49	Middle School 4f4533814b9074f6e4fb0106	Preschool 52e81612bcbc57f1066b7a45
Driving School 52e81612bcbc57f1066b7a42	High School 4bf58dd8d48988d13d941735	Music School 4f04b10d2fb6e1c99f3db0be	Private School 52e81612bcbc57f1066b7a46

סוגי מקומות הלימוד והקוד שלהם ב-API של Foursquare - https://developer.foursquare.com/docs/resources/categories

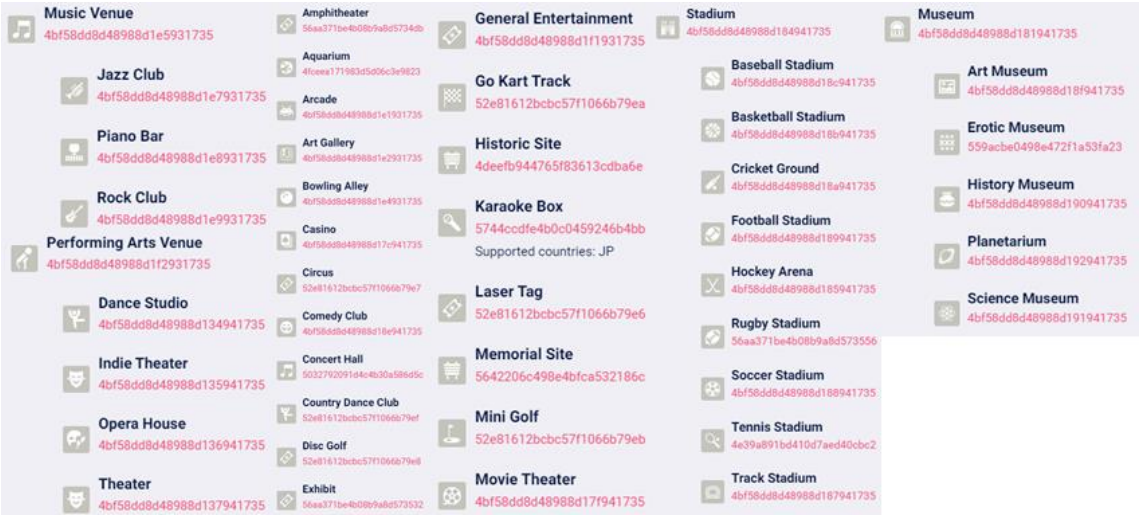




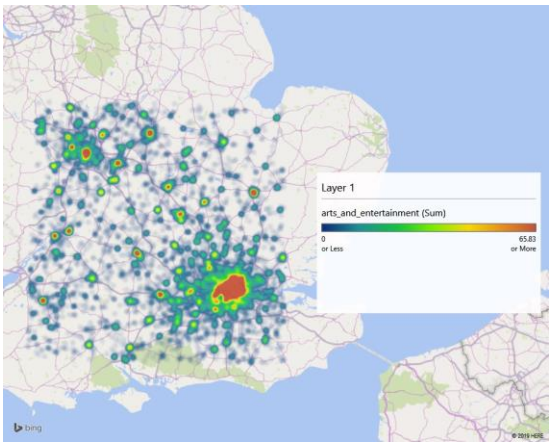
בתחום מקומות הלימוד ניתן לראות שבלונדון יש כמות גבוהה ובשאר המקומות יש פיזור קטן יותר.

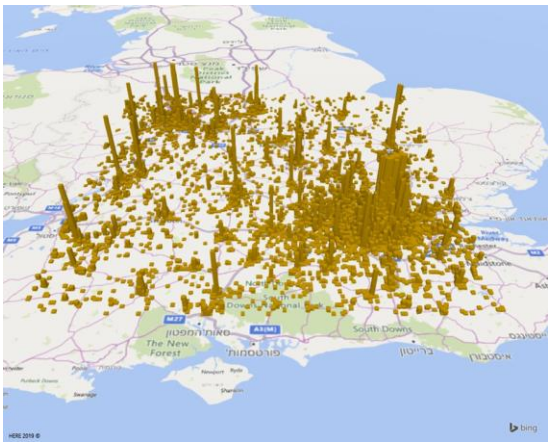
מקומות בידור ואומנות (Art & Entertainment):

סוגי מקומות בידור ואומנות שהוצאנו בעזרת ה-API של foursquare ופיזורם על המפה:

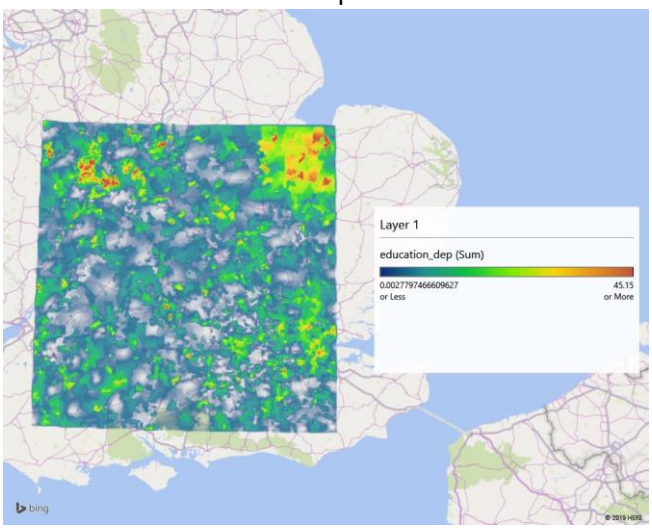


סוגי מקומות הלימוד והקוד שלהם ב-API של Foursquare - https://developer.foursquare.com/docs/resources/categories

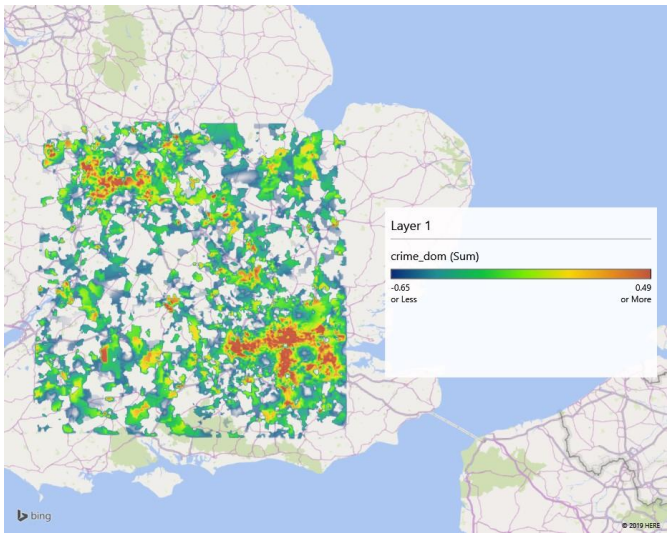




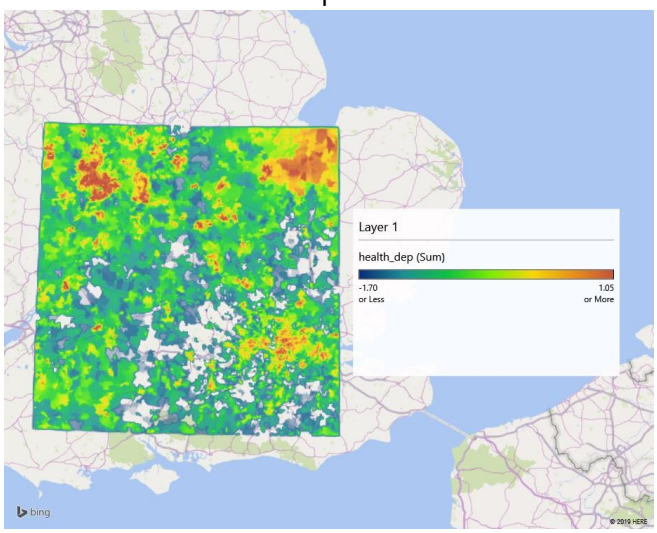
Education Deprivation



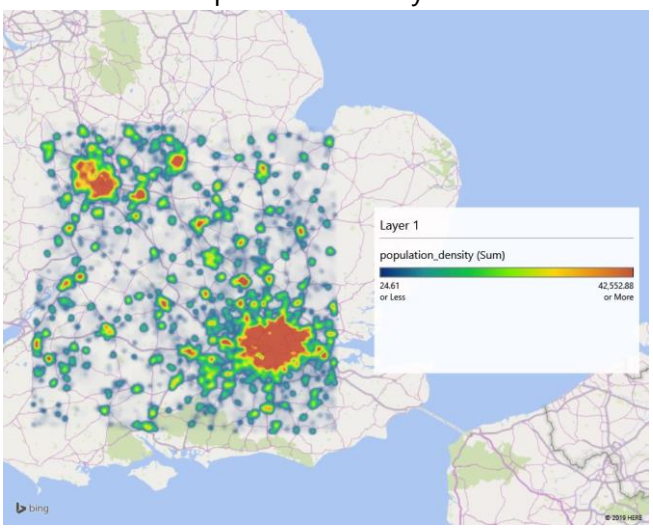
Crime Domination



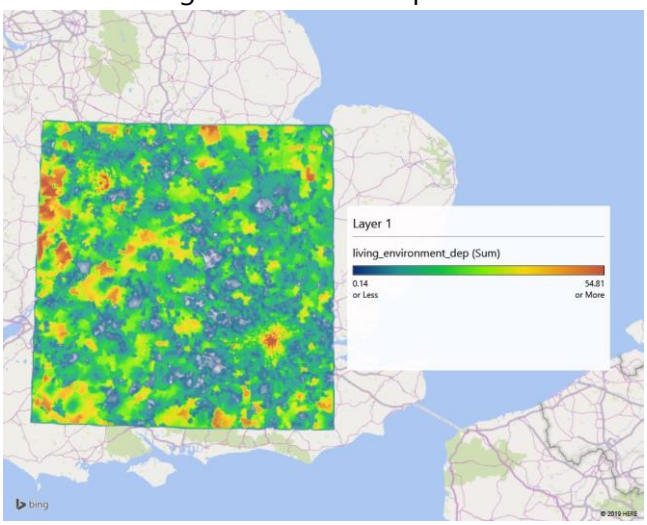
Health Deprivation



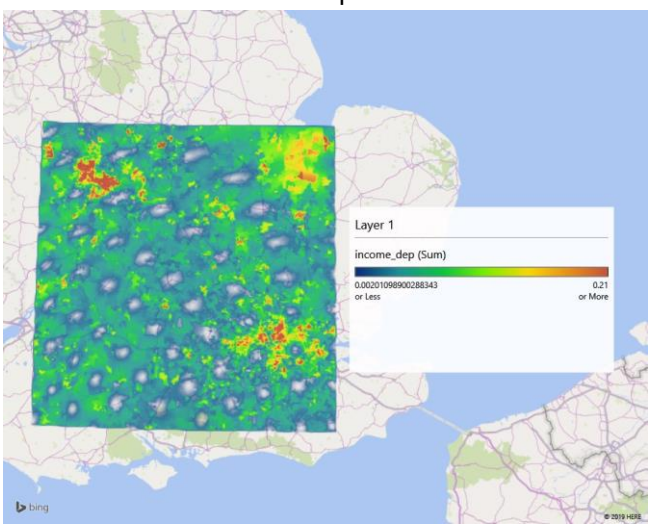
Population Density



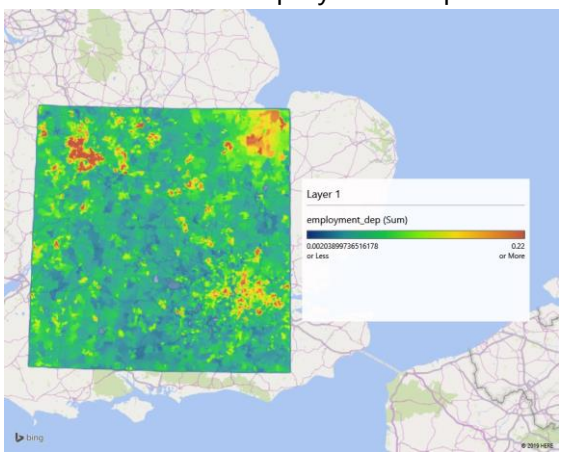
Living Environment Deprivation



Income Deprivation



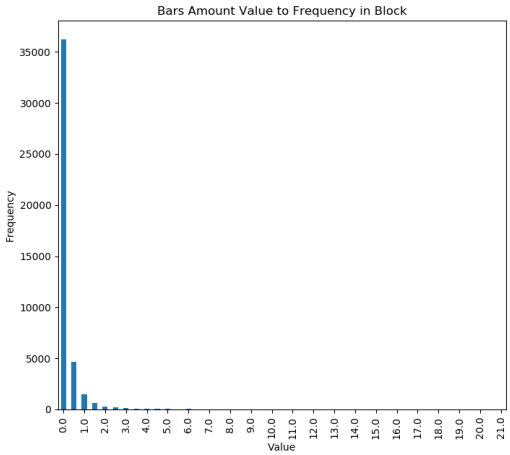
Employment Deprivation

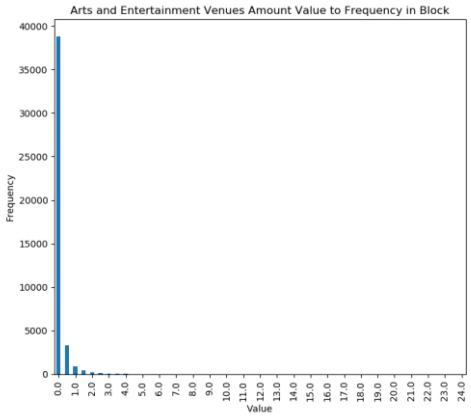


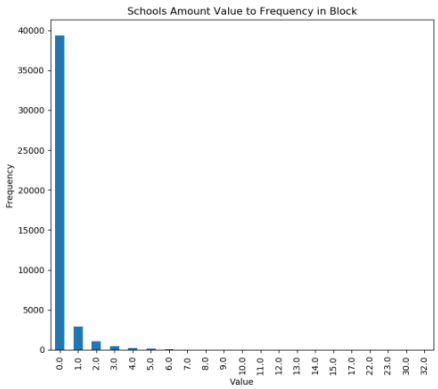
היסטוגרמות התכונות

ההיסטוגרמות הבאות מציגות לכל תכונה את כמות המופעים של כל ערך תכונה. את ההיסטוגרמות הללו הפקנו על מנת ללמוד על התפלגות התכונות שנאספו.

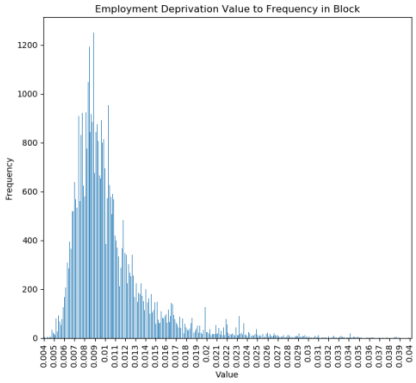
כמויות נקודות עניין-

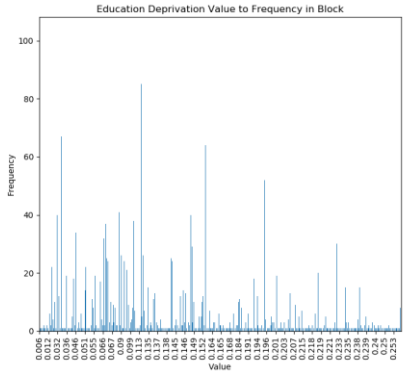


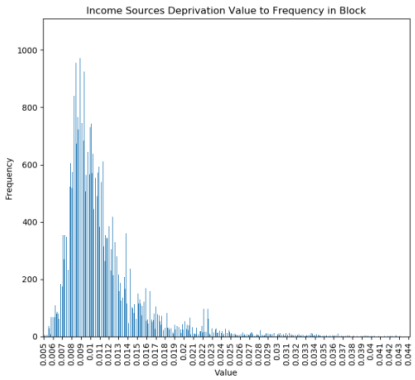


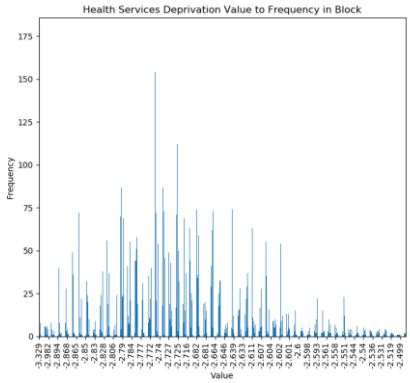


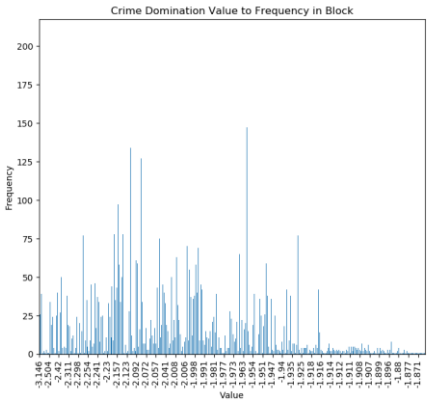
ציוני Deprivation

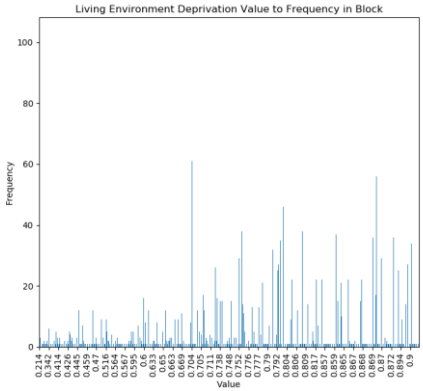












צפיפות וכמות אוכלוסייה-