Waze Challenge

Rage Against The Machine Learning

הליך עבודה- על מנת לגשת למשימה החלטנו ראשית להבין יותר לעומק את הדאטה, לשם כך פיצלנו את הדאטה למקטעי Train, Test & Dev. את הליך הPreProcessing ביצענו באופן מפוצל לשתי המשימות, אך את ניתוח הדאטה ביצענו במקביל עבור שתיהן. ראשית, התבוננו בערכים של עמודת update_data כדי לבחון את התפלגות הימים מהם מגיעות התצפיות- בשלב זה חזינו כי כל התצפיות מגיעות מ-5 תאריכים בלבד ובפרט מימים ראשון עד רביעי (**ראו נספח א'**) - כך שלא היה ברשותנו מידע על סופי שבוע וחגים. בשלב מאוחר יותר המרנו את המידע לפורמט בו אפשר לקרוא גם את השעה ביום, אז הבחנו שערכי הזמן בדידים ומופיעים בקפיצות של שעה ו-53 דקות. תובנות אלו יעזרו לנו בהמשך לבחור פיצ'רים הקשורים בזמן, ובפרט להבין שניתן להתייחס בעיקר ליום בשבוע ולשעה ביום.

ניתוח נוסף שבוצע הוא ניתוח על עמודת הערים- הבחנו כי כמחצית מהתצפיות ריקות בעמודה זו ורצינו לבחון את התנהגות המידע הריק. לשם כך בנינו קבצי SHP מקואורדינטות ה-X וה-Y, והטלנו אותם בתכנת GIS. צבענו בירוק תצפיות בהן העיר מתויגת ובשחור תצפיות בהן העיר לא מתוייגת (**נספח ב'**). ניתן לראות בבירור כי התצפיות שאינן מתוייגות הן על כבישים ראשיים, בעוד התצפיות המתויגות הן מאירועים פנים-עירוניים. מידע זה חיזק לנו את ההחלטה להשמיט מהדאטה של משימה 1 את כל התצפיות שלא מתוייגות בתל אביב, כך שלא חששנו לאבד תצפיות יקרות ערך.

משימה ראשונה- המשימה הראשונה דרשה להסתכל על רביעיית דגימות עוקבות יחד כתצפית יחידה, אשר הלייבל שלה הוא התצפית העוקבת הבאה. לכן החלוקה למקטעי עבודה Train, Test & Dev דרשה ראשית חלוקה לחמישיות ובחירה אקראית של חמישיות לכל מקטע. לאחר שהדאטה הוגרל, החלנו בתהליך הpre-processing, בו שיטחנו כל רביעיית תצפיות לשורה אחת. במקביל המרנו את הפורמטים הגולמיים לפורמטים שניתנים לעבודה- הן בתאריכים, הן עמודת linqmap_magvar אותה המרנו לסינוס וקוסינוס הזווית. בנוסף יצרנו dummies עבור מרבית המשתנים הקטגוריים, ולאחר בחינת קורלציות בין המשתנים (**נספח ג'**) החלטנו להשמיט מהמודל את הדאמיז עבור עמודות מרובות ערכים, לדוגמה עמודת הרחוב.

לאחר הPreProcess עברנו לשלב של בניית מודל Baseline – יצרנו כמה מודלים העובדים בשיטת RandomForest (גם עבור הקלסיפיקציה של Type ו-Subtype וגם עבור הרגרסיה של X ו-Y). הגענו על מקטע ה-Validation לציונים של 0.31 ב-F1-Macro על חיזוי ה-Type, 0.12 על חיזוי ה-Subtype, ומרחק של כ-20 מיליון בשגיאת המיקום. בחרנו ב-RandomForest מכיוון שרצינו להמנע מאקסטרופולציות רחוקות עבור שגיאות המיקום, וכי מנסיון ב- data challenge הוא מביא לתוצאות טובות מידית.

לאחר מכן, כתבנו סקריפט Optimize, אשר רץ על כל שילובי הHyper-parameters עבור מודלים שונים, ומשווה בין ציוני המבחן שלהם על מקטע ה-Validation. ניסינו רגרסיות וקלסיפיקציות לינאריות, RandomForest שונים עם ערכים ועומקים שונים, AdaBoost, GradientBoost, SVR, ועוד. את הבחירות הסופיות בחנו על מקטע ה-Test, על מנת שנוכל לדווח שגיאת הכללה של הבחירות שלנו. הגענו על מקטע ה-Test לציונים של 0.41 ב--F1 Macroעל חיזוי ה-Type, 0.2 על חיזוי ה-Subtype, ומרחק של כ-5 מיליון בשגיאת המיקום. (**ראו נספח ד'**)

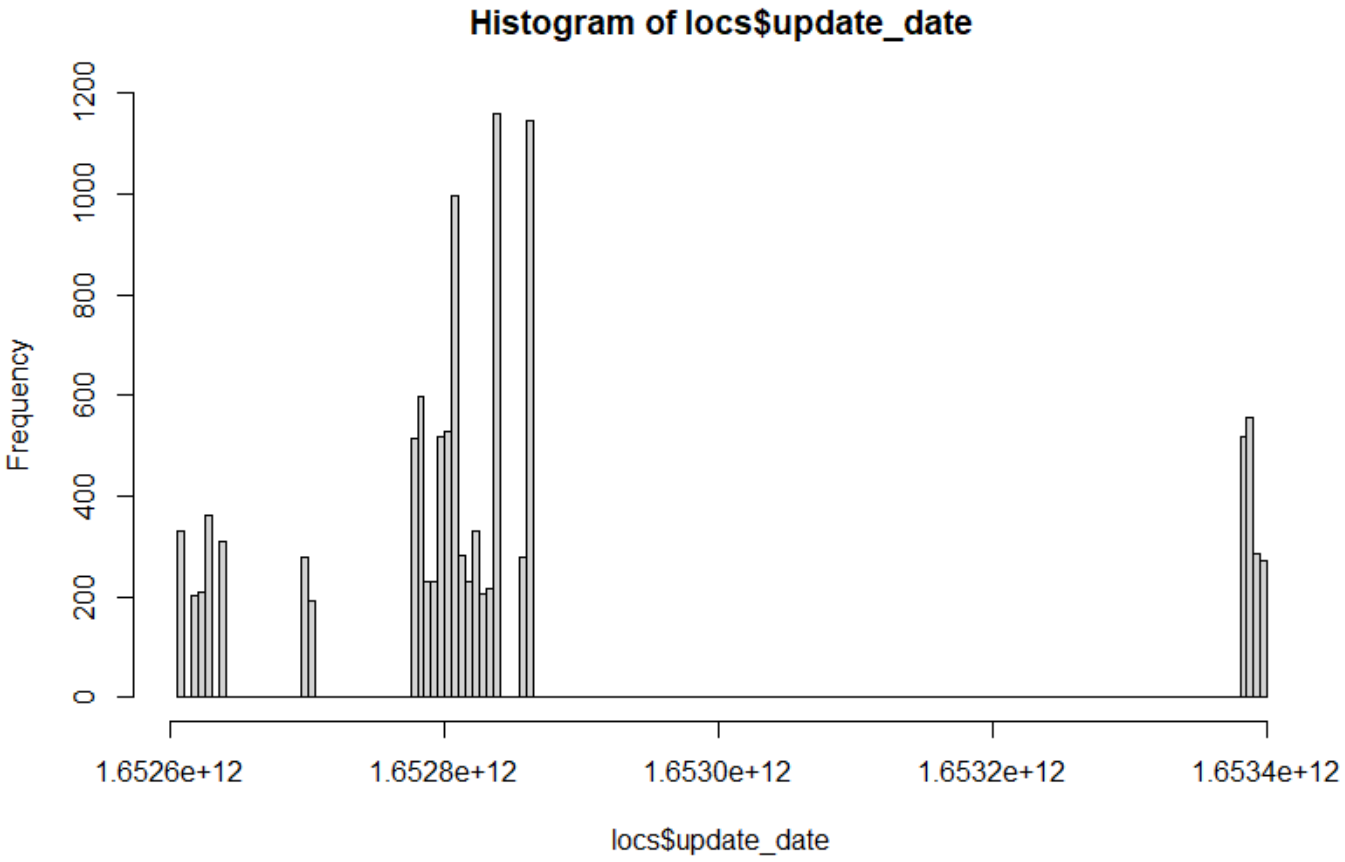
משימה שנייה- לצורך המשימה השנייה התחלנו בPreProcess דומה לראשונה מבחינת המרת פורמטים של תאריכים וחלוקה לDummies, אך לאחר מכן יצרנו טבלה חדשה בה כל תצפית מאופיינת על ידי יום בשבוע ושעה, ולכל תצפית הוספנו כפיצ'ר את מספר האירועים מכל סוג (types) שהתרחשו ביום ובשעה הנל. חלוקה זו לטבלה עזרה לנו לנתח את המידע, ובעזרת הניתוח הבחנו כי ספירת האירועים מתפלגת בצורה שנראית קרובה להתפלגות פואסון (**נספח ה'**). התפלגות זו הגיונית מאחר שהיא אכן ממדלת הסתברות שמספר מסויים של מאורעות יתרחשו בפרק זמן- באופן שתואם את המשימה.

לכן החלטנו לבנות מודל של רגרסיית פואסון- אשר יחזה בעזרת אומד MLE מה התואה הכי סבירה לכל אחד מהמאורעות. לשם כך בנינו מודל רגרסיה פואסונית מתאים לכל אחד מארבע סוגי המאורעות, ואימנו את המודל על מקטע הTrain שפיצלנו בתחילת התהליך.

לבסוף נעזרנו בעובדה שהתאריך של ה05/06/2022 הוא יום חג, ומאחר שבישראל התנועה בכבישים בימי חג ובשבתות היא כ-65% מזו שבימי חול, את התוצאות של תאריך זה הכפלנו בהתאם ב0.65.

בבדיקה בהשוואה למקטע Test, הגענו לתואצות Loss של כ400. מאחר ומדד הלוס הנ"ל זר לנו, לא היו לנו הכלים המתאימים לבחון את טיב האומד- אך כן שמנו לב, בבחינה חוזרת של הנתונים, כי החיזויים נמצאים בסדר גודל כללי קרוב לאלו של תוצאות האמת בכל אחד מהמאורעות.

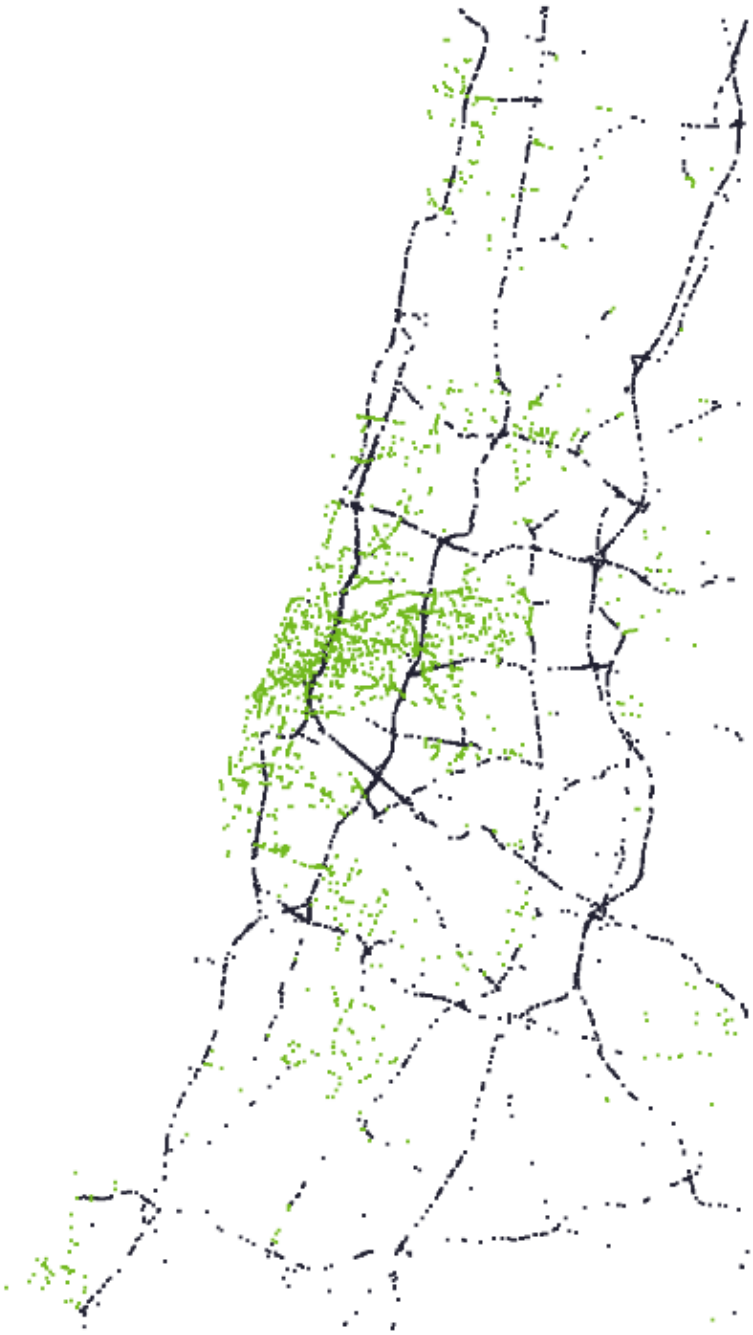
נספח א'- היסטוגרמת תאריכי תצפיות (ניתוח ראשוני)



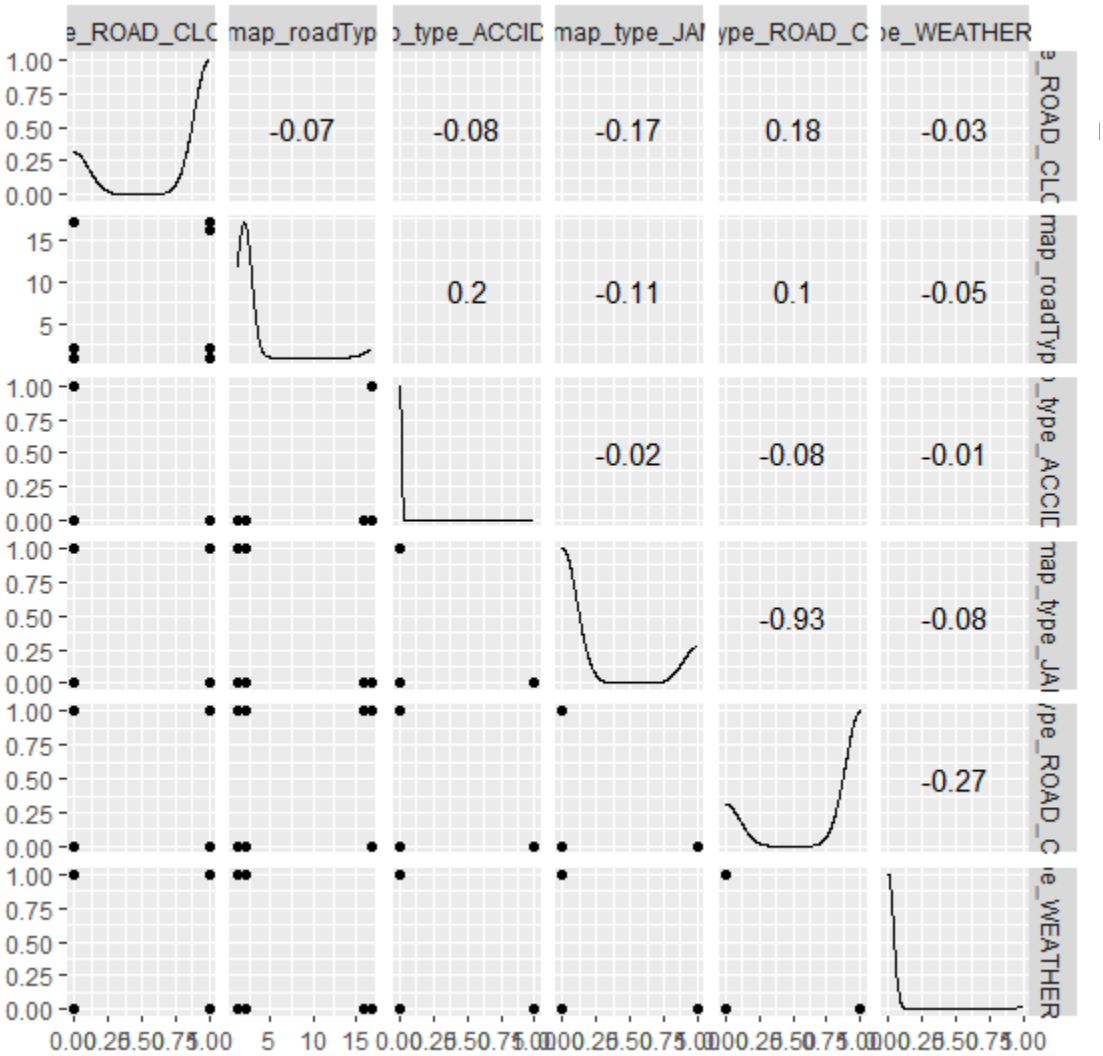
נספח ב'- קבצי שייפ של קואורדינטות, צבוע לפי תיוג עיר

תצפיות מתוייגות

תצפיות לא מתוייגות



נספח ג' – קורלציות בין (חלק) ממשתני הדאמי



נספח ד'- בחירת היפר פרמטרים עבור אחד המודלים, צבוע ע"פ לוס:

n_estimators	learning_rate	max_depth	score
90	0.3	5	0.288924
90	0.3	6	0.382705
90	0.5	3	0.374152
90	0.5	4	0.317037
90	0.5	5	0.278075
90	0.5	6	0.427477
90	0.6	3	0.324156
90	0.6	4	0.383412
90	0.6	5	0.279488
90	0.6	6	0.397181
90	0.7	3	0.41049
90	0.7	4	0.404689
90	0.7	5	0.29488
90	0.7	6	0.410343
90	0.8	3	0.370915
90	0.8	4	0.29338

נספח ה'- היסטוגרמות של צפיפות הTypes (דמויות פואסון)