התחלנו בלבחון את הדאטא. ראינו כי יש נתונים מסויימים שהנחנו שיהיה קשה לעבוד איתם, לדוגמה קיימות עמודות שערכיהן אינם מלאים, עמודות שהיו מעט מאוד ערכים שמולאו (אם בכלל) החלטנו למחוק ובעמודות שבהן היו מעט ערכים חסרים השלמנו את הערכים ע"י בדיקת הערכים הנפוצים ביותר.

הבחנו גם כי קיימים ערכים שמאתגר לעבוד איתם, לדוגמה linqmap\_magvar שמתאר את האוריינטציה ביחס לקוטב הצפוני. פיצ'ר זה מייצג זווית, ולכן זווית מאוד קטנה וזווית מאוד גדולה אמורים לייצג ערכים גדולים. הבנו כי בעצם ניתן לחשוב על הזווית כקור' x,y במישור ולכן בתהליך ה preprocessing המרנו את הזווית לשתי עמודות שהן סינוס הזווית וקוסינוס הזווית.

לאחר מכן התפצלנו לקבוצות, שניים עבדו על תהליך ה preprocessing, אחד עבד על יצירת ה “framework” להרצת המודל והאחרון עבד על חיפוש קורלציות בין נתונים בדאטא (ראו גרפים מצורפים).  
בתהליך ה preprocessing חשבנו על מספר אפשרויות, הוצאנו מהתאריך הנתון את היום בשבוע מההנחה שלנו כי "בראשון בבוקר ובחמישי בערב הפקקים הכי גרועים". כמו כן הוצאנו מהתאריך את השעה לעמודה נפרדת מההנחה כי "הפקקים משתנים בהתאם לעבודה" החלטנו לשמוט את המידע הנתון על הרחובות בהבנה כי יצירת dummies תוסיף כמות ענקית של פיצ'רים בעוד שאין כמות גדולה של דיווחים מאותם הרחובות. בנוסף איחדנו כל רבעיה של דיווחים לשורה אחת ובנוסף קיבלנו את הלייבלים מהדיווח החמישי.  
בתהליך זה התלבטנו בין שתי אופציות, אחת היא להשתמש ב “sliding window” כאשר דגימות יחזרו גם כלייבלים כדגימות עוקבות וגם כנתונים של ה train. בחרנו לא לבחור באופציה זו שכן ראינו כי קיבלנו overfit של המודל למידע האימון והחלטנו שכל נתון יופיע ברבעיה אחת בלבד של ערכים.

פיצלנו את מידע האימון ל train ו validation והתחלנו לעבוד על יצירת baseline שנוכל להשוות את תוצאות המודל הסופי שלנו אליו. ה baseline השתמש ב linear regression והשיג תוצאות של כ F1-score 0.27 בתחזית ה types כ 0.11 ב F1-score של ה subtypes ותוצאות באזור ה 10 מיליארד בשגיאת המיקום.

לאחר מספר נסיונות שונים של תחזית המידע על ידי פיצול ל multiclasses ונסיון לחזות במקביל את המידע של ה type ושל ה subtype החלטנו לעבור לתהליך דו-שלבי, בשלב הראשון נסווג את הדגימות ל type המתאים ובשלב השני נלמד מודל רק עבור הדגימות מה type המתאים על מנת לנסות לחזות את ה subtype. בנוסף על מנת לזהות את קואורדינטות המיקום של הדיווח אימנו מודל שהשתמש ב random forest regression שהשיג נתונים טובים בהרבה.