סקירה כללית של הדאטה:

קישור לדאטה: <https://www.kaggle.com/mlg-ulb/creditcardfraud>

בחרתי בדאטה סט שעוסק בעסקאות אשראי. בדאטה מפורטים פרטי עסקאות אשראי והאם העסקה היא זיוף או לא. הדאטה מציגה עסקאות שנעשו בספטמבר 2013 במשך יומיים. העסקאות מוצגות לפי הזמן שבו הם קרו ומוצגים לפי השניות שחלפו מהעסקה הראשונה. מפני שיוצרי הדאטה רצו לשמור על סודיות הם החליפו את שמות כמעט כל הפיצ'רים בשמות

V1,V2 ... V28 שאר הפיצ'רים הם Time – שמראה את הזמן שחלף (בשניות) מהעסקה הראשונה, Amount – סכום הכסף שהוציאו בעסקה, Class - 0 אם העסקה היא לא זיוף ו-1 אם היא זיוף.

דאטה לא מאוזנת:

הדאטה סט לא מאוזן יש 284,807 עסקאות שהן לא זיוף ורק 407 עסקאות מזויפות. שתי הדרכים העיקריות להתמודדות עם הבעיה הן oversampling, undersampling .

Oversampling- הוספת ערכים מהקבוצה שמופיע פחות במקרה שלנו זיוף עסקאות.

undersampling - הפחתת ערכים מהקבוצה שמופיע יותר בדאטה, במקרה שלנו עסקאות לגיטימיות (לא מזויפות).

דרכים לעשות oversampling

Random oversampling – שכפול של חלק מהמקרים מהקבוצה הקטנה (הזיופים) ושכפול שלהם. בחירת הערכים לשכפול תעשה בצורה רנדומלית.

Smote (Synthetic Minority Oversampling Technique) -