Most of the fraudulent cases are 1%, so data is unbalanced and since supervised learning is very sensitive to unbalanced data. We can use unsupervised learning:

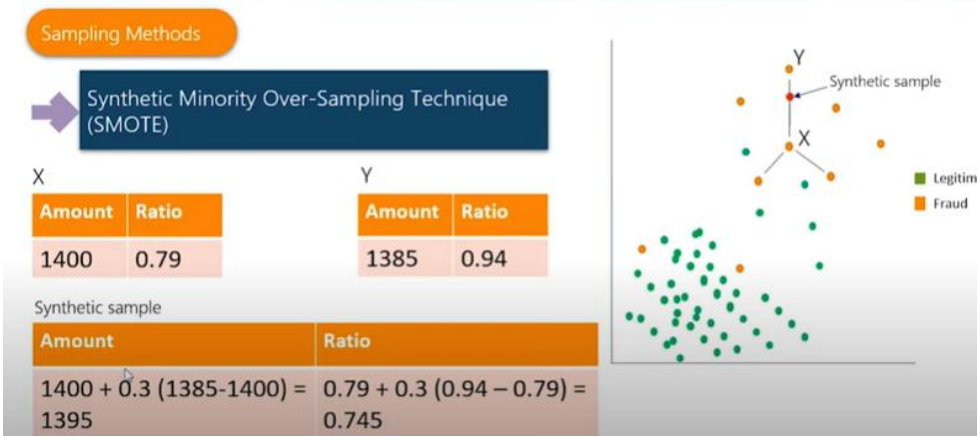
- We use PCA for reducing the features and preserving the most important features

Challenges of fraud detection

- Incorrect flagging: Avoid harassing real customers

Dealing with Unbalanced Data

- Classifier tend to favor majority class (= real/ legitimate)
- Large classification error over the fraud cases
- Classifiers learn better from a balanced distribution
- We will balance the training set
    - 1st technique is random over sampling: We copy the lesser observations (here fraudulent class) until we reach a threshold value. Problem is there will be lots of observations what we get after copying and there will be variance which can't be explained by the data
    - Random under sampling: We remove the dominant cases (legitimate cases). Problem is we will throw lots of useful information that is not preferred in general
    - We can do both over-sampling and under-sampling
    - SMOTE (synthetic minority over sampling technique): In this technique, we over-sample the fraud cases (minority class) by creating synthetic fraud cases.



- Find k nearest neighbour of fraud case (suppose k=3), randomly choose one neighbour say Y, now we have to choose any point on line between x and y.
- We will do this repeatedly to get different synthetic fraudulent case points