Dealing with imbalanced datasets

Mistakes to correct:

* Never test on over/under sampled datasets
* If we use CV oversample or undersample during it otherwise we will be subject to a data leakage problem
  + The issue is by over or under sampling we indirectly influence the validation set
  + The idea is we have original dataset, we split via cross validation then oversample for the training set
* Don’t use accuracy use f1/precision/recall instead

What they do:

* EDA
* Use RobustScaler from sklearn.preprocessing to scale appropriate columns
* Use StratifiedKFold and loop over these folds which contain an equal number of positive cases
  + Check if train and test labels are similarly distributed (look at proportions)
* Implements **Random Under-Sampling** (remove data to have a more balanced dataset)
  + Df.sample(frac = 1) to shuffle
  + Get fraud\_df and non\_fraud\_df where the latter is the same size at the former
  + Combine df
  + Shuffle rows again in the new df
* Finds **correlation matrices** and plots box-plots
* **Anomaly detection** – want to remove extreme outliers from features which have high correlation (improves accuracy of model)
  + IQR method – remove all values outside the 25th -75th percentile range
    - Usually multiply by some factor say 1.5 by the IQR changing the number of outliers detected
    - Only looks at cases of outliers where given the case of fraud the distribution is normal
* Dimensionality Reduction and Clustering:
  + Uses **t-SNE** algorithm for clustering
    - Uses T-SNE, PCA and truncated SVD
    - Does a scatter plot and observes clusters
* Splits into test and train (from sub-sample), implement 4 classifiers Logistic, KNeighborsClassifier, SVC and DecisionTreeClassifier
  + Uses GridSearchCV to find optimal parameters
* Implements NearMiss as a distribution check
  + NearMiss is an undersampling technique which selects points closest to the distribution of the positive cases
* Note the above was done incorrectly since CV was implemented on the undersampled object, the notebook then uses stratifiedKFold
  + Notebook then does

A computer screen with white text

AI-generated content may be incorrect.

* + Plots the learning curve for each estimator
  + Finds AUC score
  + Plots ROC curve
* Goes through some properties of Logistic regression
* Then goes through **SMOTE** 
  + Picks distances between neighbours of minority class
  + More accurate but also greater computation