**What is data imbalance?**

Imbalanced data or classes is a common problem in machine learning classification where there are a disproportionate ratio of observations in each class.

**Why data imbalance?**

Data imbalance occurs when you collect disproportionate ratio of data, thus one class of data is much more than the other.

It is important to balance data because of biasness in predicting the output, since the machine will learn more from one class making it have very high probability of classifying the output as the class with higher proportion.

**What data are we balancing?**

Data that contains unequal proportion of the classes of data to be used in training the model.

**Methods used to balance data.**

1.Change the performance metric

Accuracy is not the best metric to use when evaluating imbalanced datasets as it can be very misleading. Metrics that can provide better insight include:

Confusion matrix

Precision

Recall

F1 score

**2. Resampling Techniques – Oversample minority class**

Oversampling can be defined as adding more copies of the minority class. Oversampling can be a good choice when you don’t have a ton of data to work with. Always split into test and train sets BEFORE trying oversampling techniques! Oversampling before splitting the data can allow the exact same observations to be present in both the test and train sets. This can allow our model to simply memorize specific data points and cause overfitting and poor generalization to the test data.

**3. Change the algorithm**

While in every machine learning problem, it’s a good rule of thumb to try a variety of algorithms, it can be especially beneficial with imbalanced datasets. Decision trees frequently perform well on imbalanced data. They work by learning a hierarchy of if/else questions and this can force both classes to be addressed.

**4. Resampling techniques — Undersample majority class**

Undersampling can be defined as removing some observations of the majority class. Undersampling can be a good choice when you have a ton of data -think millions of rows. But a drawback is that we are removing information that may be valuable. This could lead to underfitting and poor generalization to the test set.

**5. Generate synthetic samples**

A technique similar to upsampling is to create synthetic samples. Here we will use imblearn’s SMOTE or Synthetic Minority Oversampling Technique. SMOTE uses a nearest neighbors algorithm to generate new and synthetic data we can use for training our model. Again, it’s important to generate the new samples only in the training set to ensure our model generalizes well to unseen data.