Handling Imbalanced Datasets

Hager Zeyada

February 2022

Abstract

A dataset is imbalanced if the classification categories are not approximately equally represented. Recent years brought increased interest in applying machine learning techniques to difficult "real world" problems, many of which are characterized by imbalanced data. Additionally the distribution of the testing data may differ from that of the training data, and the true misclassification costs may be unknown at learning time. Predictive accuracy, a popular choice for evaluating performance of a classifier, might not be appropriate when the data is imbalanced and/or the costs of different errors vary markedly.

1 Introduction

High imbalance occurs in real world domains where the decision system is aimed to detect a rare but important case. A number of solutions to the class imbalance problem were proposed at both the data and algorithmic levels. At the data level, these solutions include many different forms of resampling, such as random oversampling with replacement (adding more copies to the minority class. It can be a good choice when you don't have a ton of data to work with), random undersampling (removing some observations of the majority class. This is done until the majority and minority class is balanced out), directed oversampling (in which no new examples are created, but the choice of samples to replace is informed rather than random), directed undersampling (where the choice of examples to eliminate is informed), or even combinations of the above techniques. [3]

As for the algorithmic level, solutions include adjusting the costs of the various classes, in a way to counter the class imbalance, which is known as Penalize Algorithm (Cost-Sensitive Training), adjusting the probabilistic estimate at the tree leaf (when working with decision trees), adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning. [1]

Typical ways of dealing with this issue are resampling the dataset (either oversampling the minority classes or undersampling the majority classes. In this paper, you will discuss the approaches of dealing with imbalanced datasets, based on a mixture of supervised and unsupervised learning.

Feature	Datatype
age	int64
sex	int64
cp	int64
trtbps	int64
chol	int64
fbs	int64
restecg	int64
thalachh	int64
exng	int64
oldpeak	float64
slp	int64
caa	int64
thall	int64
output	int64

Table 1: Heart Attack Analysis and Prediction Dataset Features Datatypes.

2 Data

2.1 Heart Attack Analysis and Prediction Dataset

Heart Attack Analysis and Prediction Dataset, is a balanced dataset collected from Kaggle with shape (303,14), which has data on patients seen by a cardiologist.The problem type of this dataset is "Classification", as the main goal is to build a machine learning model, that will be able to predict the risk of a heart attack based on a patient's health condition as "0" or "1", see Table 2. Table 1 represents the dataset features along with the datatype of each feature. This dataset is suitable for this project as it is a balanced dataset (54.3% of one class), see Figure1. A pre-processing is made to the data so we remove the null values and duplicates if exists. As this paper is going to discuss how we can handle the imbalanced datasets, I created three versions of the dataset, with different imbalances in one class. The new datasets will be with low, medium and high imbalances (65%, 75%, 90%) respectively. Then a function is applied to check the balance of every new dataset, see Figure2, Figure3, Figure4

2.2 Airline Passenger Satisfaction Dataset

Airline Passenger Satisfaction Dataset, is a balanced dataset collected from Kaggle with shape (103904, 25). This dataset contains an airline passenger satisfaction survey. The target column in this dataset is "Satisfaction" which indicates whether a passenger is satisfied or dissatisfied, see Table 3. Table 4 represents the dataset features along with the datatype of each feature. This dataset is suitable for this project as it is a balanced dataset (56.6% of one class), see Figure5. A pre-processing is made to the data so we remove the null values and duplicates if exists. Three versions of the dataset were originated

age	age in years
sex	sex (0 = female; 1 = male)
cp	chest pain type (1 = typical angina; 2 = atypical angina;3 = non-anginal pain; 0 = asymptomatic)
trtbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	fasting blood sugar ¿ 120 mg/dl (0 = false; 1 = true)
restecg	resting electrocardiographic results (0 = normal; 1 = hypertrophy; 2 = having ST-T wave abnormality)
thalachh	maximum heart rate achieved
exng	exercise induced angina (0 = no; 1 = yes)
oldpeak	ST depression induced by exercise relative to rest
slp	the slope of the peak exercise ST segment (0 = downsloping; 1 = flat; 2 = upsloping)
caa	number of major vessels (0-4) colored by flourosopy
thall	thallium stress test (1 = fixed defect; 2 = reversable defect; 3 = normal)
output	0 = less chance of heart attack; 1 = more chance of heart attack

Table 2: Heart Attack Analysis and Prediction Dataset Features.

from the dataset, with different imbalances in one class. The new datasets will be with low, medium and high imbalances (65%, 75%, 90%) respectively. Then a function is applied to check the balance of every new dataset, see Figure6, Figure7, Figure8

2.3 Raisin Dataset

Raisin Dataset, is a balanced dataset collected from UCI with shape (900, 8). The dataset was originally formed by collecting images of Kecimen and Besni raisin varieties grown in Turkey were obtained with CVS. A total of 900 raisin grains were used, including 450 pieces from both varieties. These images were subjected to various stages of pre-processing and 7 morphological features were extracted. These features have been classified using three different artificial intelligence techniques. The target column in this dataset is "Class", in which we classify the raisin into "Kecimen" or "Besni", see Table 5. Table 6 represents the dataset features along with the datatype of each feature. This dataset is suitable for this project as it is a balanced dataset (50.0% of each class), see Figure9. A pre-processing is made to the data so we remove the null values and duplicates if exists. A three versions of the datasets were created, with different imbalances in one class. The new datasets will be with low, medium and high imbalances (65%, 75%, 90%) respectively. Then a function is applied to check the balance of every new dataset, see Figure10, Figure11, Figure12

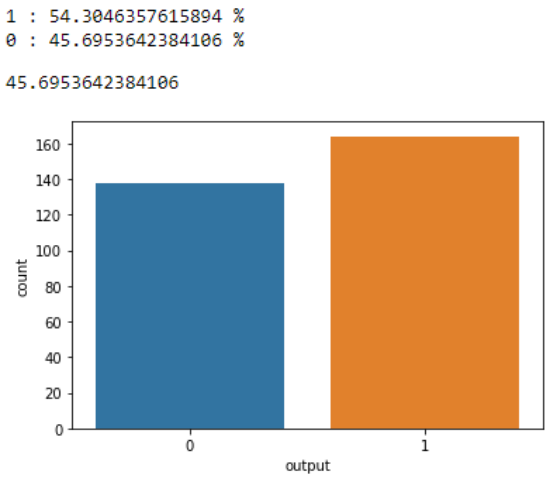

Figure 1: Heart Attack Analysis and Prediction Balanced Dataset.

3 Methodology

In this project, we will analyze and test a new approach of dealing with imbalanced datasets, based on a mixture of supervised and unsupervised learning methods, in which we can train the our model and correctly predict unseen data.

As mentioned before, there are two main approaches to make a balanced dataset out of an imbalanced one, which are under-sampling and oversampling. I used undersampling, so I can generate the three versions of imbalances of each dataset.

I used the undersampling techinque to create the new versions of the imbalances of each dataset. To start working on the datsets, we will establish a baseline, and perform stratified cross-validation on each of the datasets and their surrogates and train a random forest.

Then we will perform a stratified 10-Folds cross-validation, so that we provide train/test indices to split data in train/test sets. This cross-validation object is a variation of the K-Folds that returns stratified folds, and the folds are made by preserving the percentage of samples for each class.

We will use the data of 9 of these folds, and perform the Elbow method and the Silhouette method, to determine the optimum number of clusters in the datasets.

Then we will run K-means in the dataset using the identified number of clusters. For each cluster, we will identify the cluster centroid and the number of samples of the minority class in that cluster. A random forest will be trained for each of the clusters that contains samples from more than one class. For each sample from the unseen fold, we will assign each sample to its closest cluster.

This will be repeated for each permutation of the 10 folds. Finally, this method will be compared to the baseline model using the evaluation metric F-

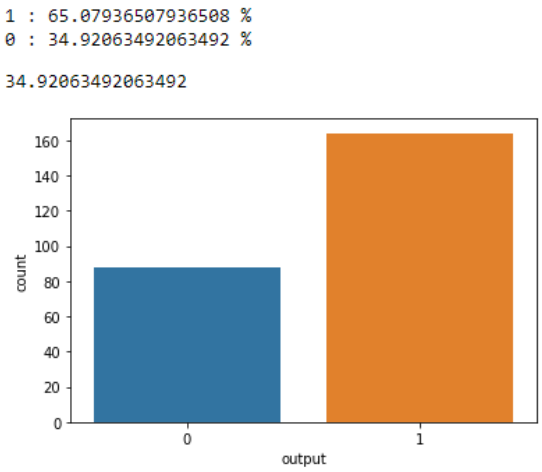

Figure 2: Heart Attack Analysis and Prediction Low Imbalance (65%) Dataset.

Score to determine if this approach is better for the classification of imbalanced datasets.

4 Conclusion

It is often reported that cost-sensitive learning outperforms random re-sampling [4]. The relationship between training set size and improper classification performance for imbalanced data sets seems to be that on small imbalanced data sets the minority class is poorly represented by an excessively reduced number of examples that might not be sufficient for learning, especially when a large degree of class overlapping exists and the class is further divided into sub-clusters. On contrary, larger data sets, the effect of these complicating factors seems to be reduced, as the minority class is better represented by a larger number of examples, which will be more sufficient for learning.

5 Code

https://github.com/HagarMostafa/CE888/tree/main/project

References

[1] www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/

[2] arxiv.org/pdf/1106.1813.pdf

```
1 : 75.22935779816514 %
0 : 24.770642201834864 %

24.770642201834864
```

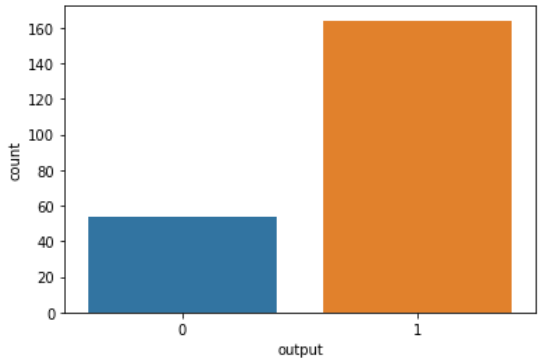

Figure 3: Heart Attack Analysis and Prediction Medium Imbalance (75%) Dataset.

[3] www.researchgate.net/publication/228084509_Handling_imbalanced_datasets_A_review

[4] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. Intelli-gent Data Analysis, 6(5):203-231, 2002.

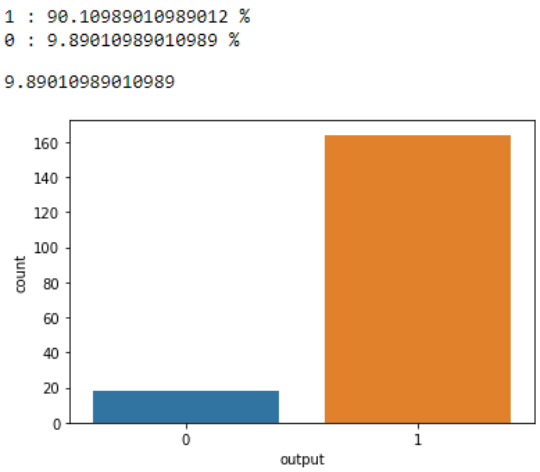

Figure 4: Heart Attack Analysis and Prediction High Imbalance (90%) Dataset.

Gender	Gender of the passengers (Female, Male)
Customer Type	The customer type (Loyal customer, disloyal customer)
Age	The actual age of the passengers
Type of Travel	Purpose of the flight of the passengers (Personal Travel, Business Travel)
Class	Travel class in the plane of the passengers (Business, Eco, Eco Plus)
Flight distance	The flight distance of this journey
Inflight wifi service	Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
Departure/Arrival time convenient	Satisfaction level of Departure/Arrival time convenient
Ease of Online booking	Satisfaction level of online booking
Gate location	Satisfaction level of Gate location
Food and drink	Satisfaction level of Food and drink
Online boarding	Satisfaction level of online boarding
Seat comfort	Satisfaction level of Seat comfort
Inflight entertainment	Satisfaction level of inflight entertainment
On-board service	Satisfaction level of On-board service
Leg room service	Satisfaction level of Leg room service
Baggage handling	Satisfaction level of baggage handling
Check-in service	Satisfaction level of Check-in service
Inflight service	Satisfaction level of inflight service
Cleanliness	Satisfaction level of Cleanliness
Departure Delay in Minutes	Minutes delayed when departure
Arrival Delay in Minutes	Minutes delayed when Arrival
Satisfaction	Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

Table 3: Airline Passenger Satisfaction Dataset Features.

Feature	Datatype
id	int64
Gender	object
Customer Type	object
Age	int64
Type of Travel	object
Class	object
Flight Distance	int64
Inflight wifi service	int64
Departure/Arrival time convenient	int64
Ease of Online booking	int64
Gate location	int64
Food and drink	int64
Online boarding	int64
Seat comfort	int64
Inflight entertainment	int64
On-board service	int64
Leg room service	int64
Baggage handling	int64
Checkin service	int64
Inflight service	int64
Cleanliness	int64
Departure Delay in Minutes	int64
Arrival Delay in Minutes	float64
satisfaction	object

Table 4: Airline Passenger Satisfaction Dataset Features Datatypes.

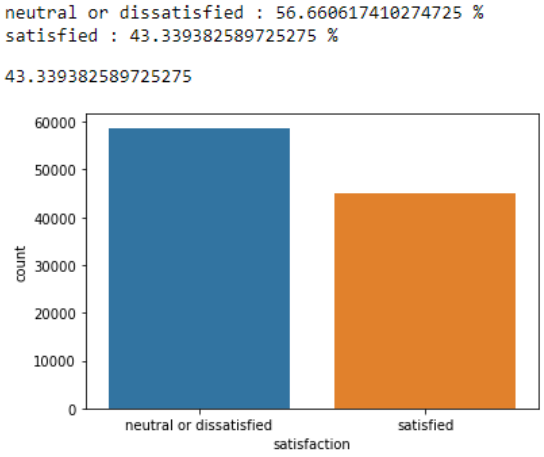

Figure 5: Airline Passenger Satisfaction Balanced Dataset.

```
neutral or dissatisfied : 65.09592991016969 %
satisfied : 34.90407008983032 %

34.90407008983032
```

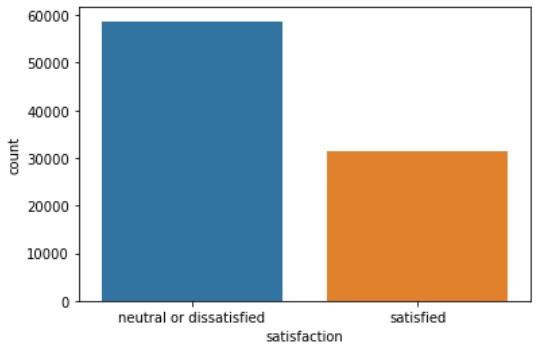

Figure 6: Airline Passenger Satisfaction low Imbalance (65%) Dataset.

```
neutral or dissatisfied : 75.03323618141842 %
satisfied : 24.966763818581583 %

24.966763818581583
```

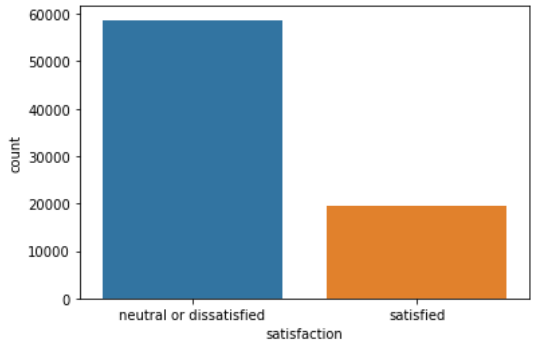

Figure 7: Airline Passenger Satisfaction Medium Imbalance (75%) Dataset.

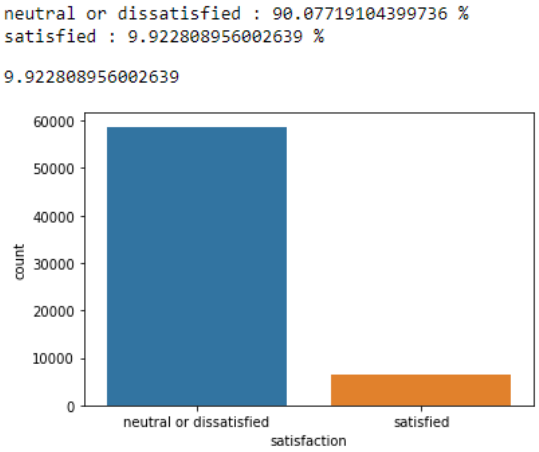

Figure 8: Airline Passenger Satisfaction High Imbalance (90%) Dataset.

Area	Gives the number of pixels within the boundaries of the raisin.
Perimeter	It measures the environment by calculating the distance between the boundaries of the raisin and the pixels around it.
MajorAxisLength	Gives the length of the main axis, which is the longest line that can be drawn on the raisin.
MinorAxisLength	Gives the length of the small axis, which is the shortest line that can be drawn on the raisin.
Eccentricity	It gives a measure of the eccentricity of the ellipse, which has the same moments as raisins.
ConvexArea	Gives the number of pixels of the smallest convex shell of the region formed by the raisin.
Extent	Gives the ratio of the region formed by the raisin to the total pixels in the bounding box.
Class	Kecimen and Besni raisin.

Table 5: Raisin Dataset Features.

Feature	Datatype
Area	int64
MajorAxisLength	float64
MinorAxisLength	float64
Eccentricity	float64
ConvexArea	int64
Extent	float64
Perimeter	float64
Class	object

Table 6: Raisin Dataset Features Datatypes.

```
Kecimen : 50.0 %
Besni : 50.0 %

50.0
```

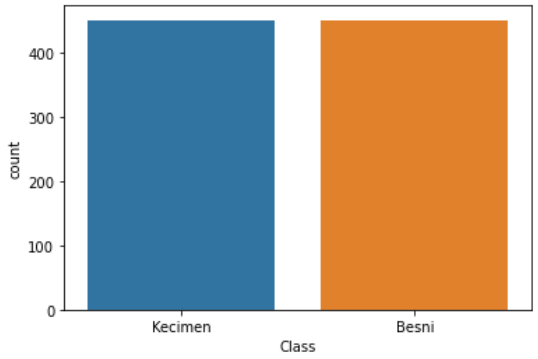

Figure 9: Raisin Balanced Dataset.

```
Kecimen : 34.971098265895954 %
Besni : 65.02890173410405 %

65.02890173410405
```

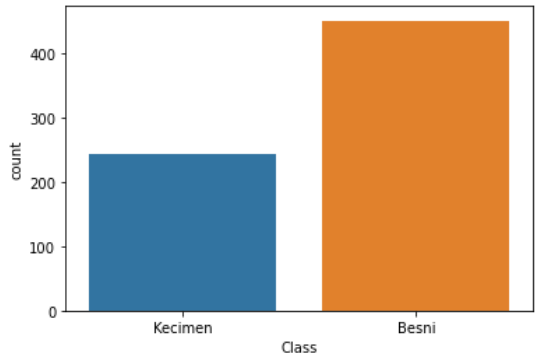

Figure 10: Raisin low Imbalance (65%) Dataset.

```
Kecimen : 25.0 %
Besni : 75.0 %

75.0
```

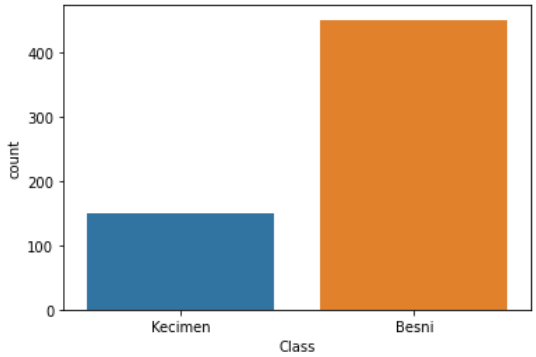

Figure 11: Raisin Medium Imbalance (75%) Dataset.

```
Kecimen : 10.0 %
Besni : 90.0 %

90.0
```

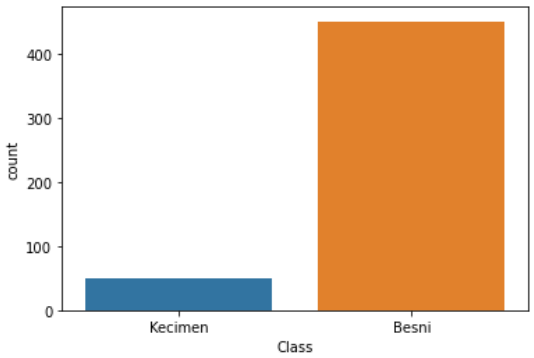

Figure 12: Raisin High Imbalance (90%) Dataset.