Class Imbalance, SMOTE, borderline SMOTE, ADASYN

Class Imbalance can put our algorithm off balance

Dr. Saptarsi Goswami · Follow

Published in Towards Data Science
5 min read · Nov 2, 2020

Listen Share More

Image Source

It was the late 1990s when Niesh V Chawla (The main brain behind SMOTE), then a graduate student at the University of South Florida was working on a binary classification problem. He was dealing with mammography images and his task was to build a classifier that will take a pixel as an input and will classify it as a normal pixel or a cancerous pixel. He was quite happy when he achieved 97% classification accuracy. His happiness was shortlived when he saw 97.6% of pixels were normal.

You may be thinking, what's the problem? There are two issues

- Let's say in a sample of 100 pixels, 98 are normal and 2 are cancerous, and if we write a program, which predicts anything to be normal. What will be the classification accuracy? A whopping 98%. **Did the program learn? Not at all.**
- There is yet another issue. The classifiers strive to get good performance in the training data and as normal observations are more, **they will concentrate on learning the pattern of the 'normal' class more.** It's just like what any student would do when they know 98% of questions will come from Algebra and 2% from Trigonometry. They will safely ignore Trignometry

So, why this problem is manifested it's because there is a great disparity between the frequency or count of the classes. **We call such a dataset to exhibit class imbalance. The normal class is referred to as the majority class and the rare class is called the minority class.**

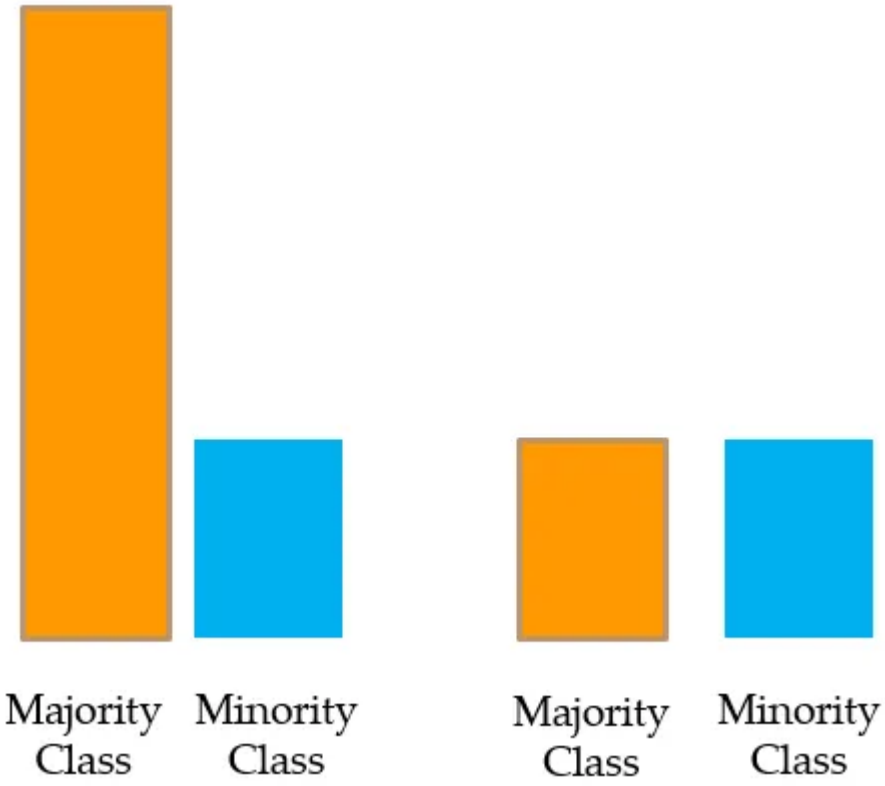
A Seagull as the minority class (https://unsplash.com/photos/SQBtDa6cKDA)

Does this exist in real-life applications? Take the example of spam detection, fake news detection, fraud detection, suspicious activity detection, intrusion detection, etc, where the class imbalance problem is manifested.

Solutions to bringing some balance:

The basic approaches are called resampling techniques. There are two basic approaches.

Undersampling:-

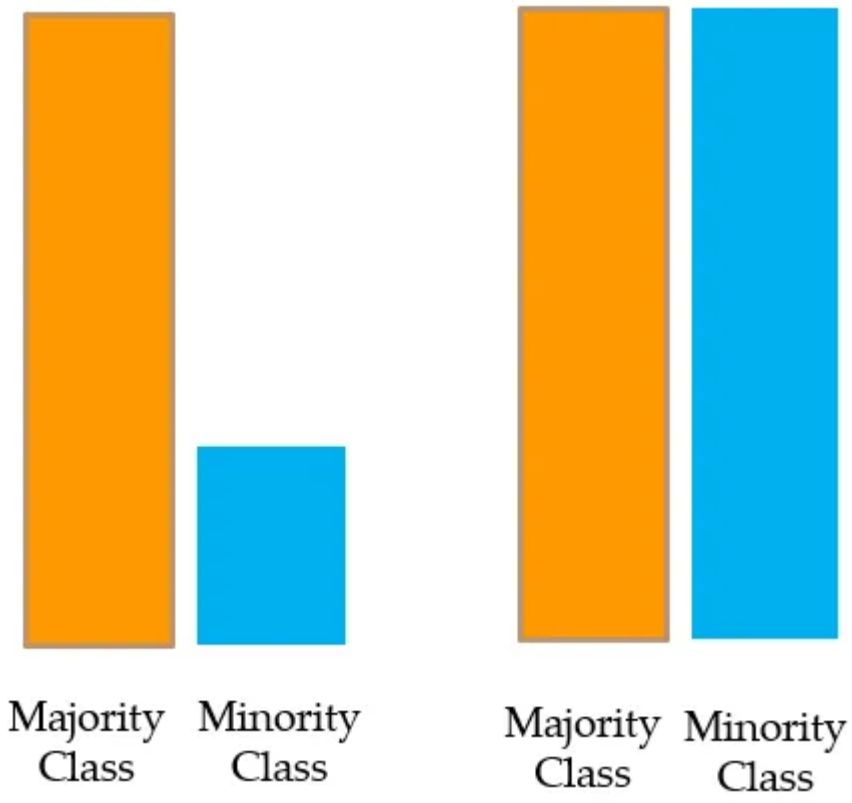


Undersampling or downsampling the majority class (Image Source Author)

We pick random samples from the majority class and make it equal to the minority class count. This is called the **undersampling or downsampling of the majority class.**

Issue: **It's not a good idea to ignore or let go so much of original data.**

Oversampling:-



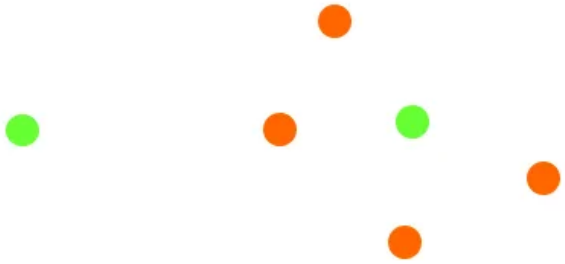
Oversampling or upsampling the minority class (Image Source Author)

Here, sampling with replacement is applied to the minority class to create as many observations as there are in the majority class and the two classes are balanced. This is called **oversampling or upsampling** the minority class.

Issue: **Repetition of the same minority class data leads to overfitting.**

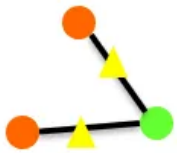
SMOTE:

The full form of SMOTE, **S**ynthetic **M**inority **O**versampling **Te**chnique. Here Synthetic observations are generated from the Minority class



Step 1:
Identify a point from the minority class

Step 2:
Identify it's nearest neighbors. (Here 4 neighbors are taken. All minority neighbors)



Step 3:
Assume we want to add 2 times more synthetic observations
Select 2 nearest neighbor randomly
Draw a line between the point and the selected neighbors
Create a synthetic observation along the line

SMOTE, Synthetic Minority Observation Generation Process (Source: Author)

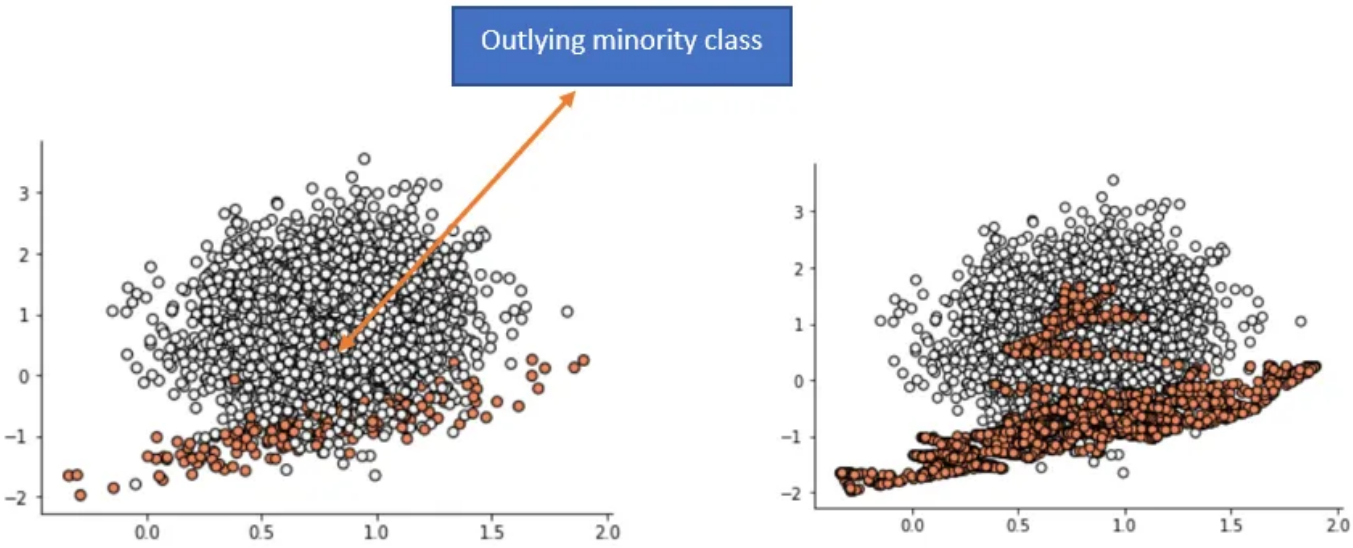
Let there be two observations (x1,y1) and (x2,y2) from the minority class. As a first step, a random number between 0 and 1 is created, let's call it r. The synthetic point will be (x1 + r*(x2 -x1), y1 + r*(y2 -y1)). It's illustrated further with the following example.

Original Data Points	
Attr 1	Attr 2
6	4
4	3

Synthetic Data Points		
r	Attr 1	Attr 2
0.2	5.6	3.8
0.5	5	3.5
0.8	4.4	3.2

Synthetic points generate from minority class (Image Source: Author)

An issue with SMOTE:

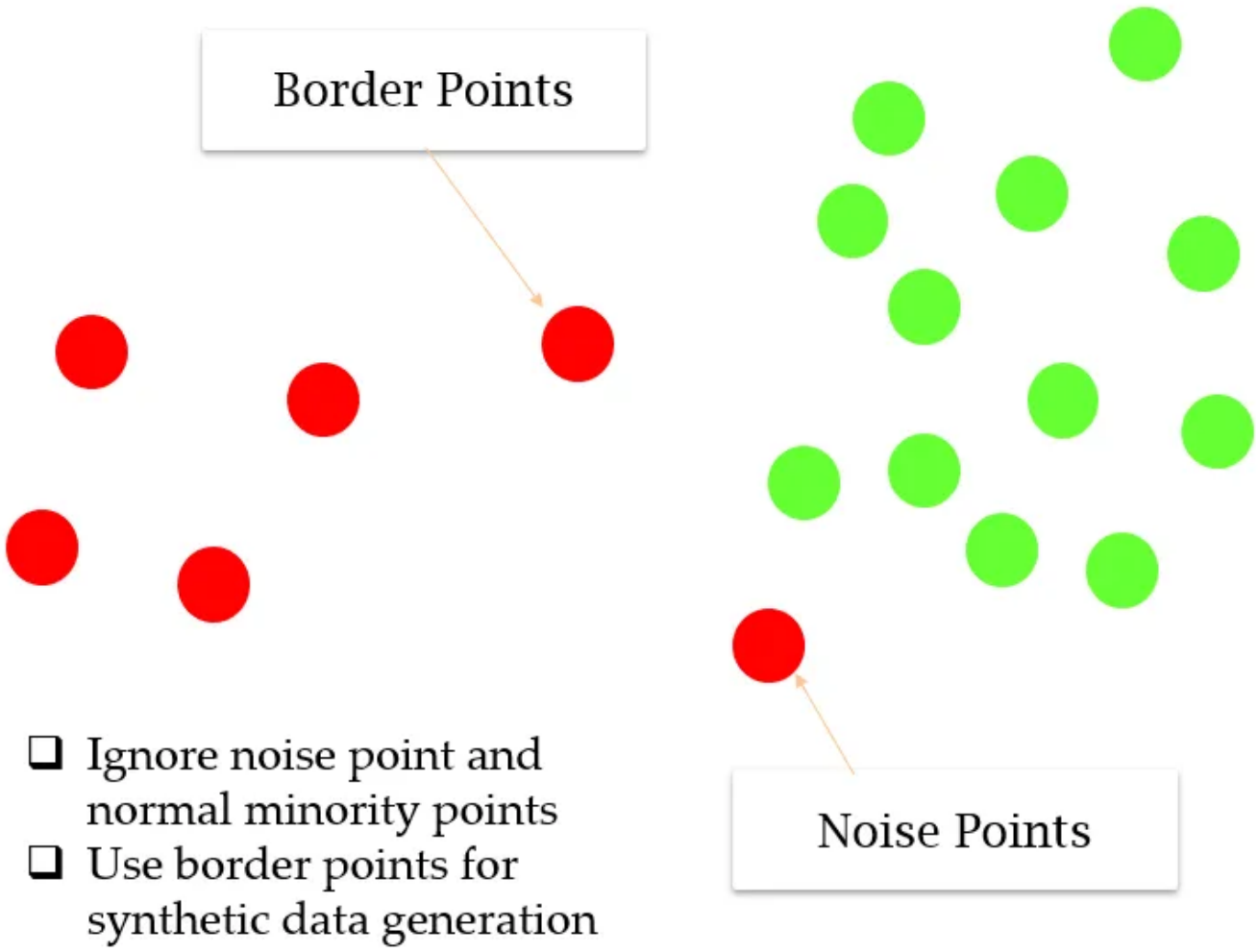


Left Side: Original Data Right Side: Data after SMOTE is Applied (Image Source: Author)

If there are observations in the minority class which are outlying and appears in the majority class, it causes a problem for SMOTE, by creating a line bridge with the majority class.

Borderline SMOTE:-

This solves the above issue.



Border Line SMOTE : (Image Source Author)

This algorithm starts by classifying the minority class observations. It classifies any minority observation as a noise point if all the neighbors are the majority class and such an observation is ignored while creating synthetic data (Similar to DBSCAN). Further, it classifies a few points as border points that have both majority and minority class as neighborhood and resample completely from these points (Extreme observations on which a support vector will typically pay attention to).

Issue: End up giving more attention to these extreme observations.

ADASYN:

ADASYN is a more generic framework, for each of the minority observations it first finds the impurity of the neighborhood, by taking the ratio of majority observations in the neighborhood and k.

Minority Class	Minority Neighbours	Majority Neighbours	Impurity Ratio
Obs 1	3	2	.6
Obs 2	4	1	.4
Obs 3	1	4	.8
Obs 4	5	0	0

ADASYN Impurity Ratio

Now, first of all, this impurity ratio is converted into a probability distribution by making the sum as 1. Then higher the ratio more synthetic points are generated for that particular point. **Hence the number of synthetic observations to be created for Obs 3 is going to be double that of Obs 2.** So it's not so extreme as Borderline SMOTE and the boundary between the noise point, border point, and regular minority points are much softer. (Not a hard boundary). Thus the name adaptive.

This is also explained in the following video tutorial, please give it a watch

Endnote:

Class Imbalance is a very practical problem. Resampling-based approaches not promising, which motivated researchers to develop SMOTE, which was gradually improved by borderline SMOTE, ADASYN, etc. For learning about more of the variants reference 2 is a great read.

Reference

[1] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002 Jun 1;16:321–57.

[2] Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research. 2018 Apr 20;61:863–905.

Class Imbalance | Smote | Imbalanced Class | Data Science

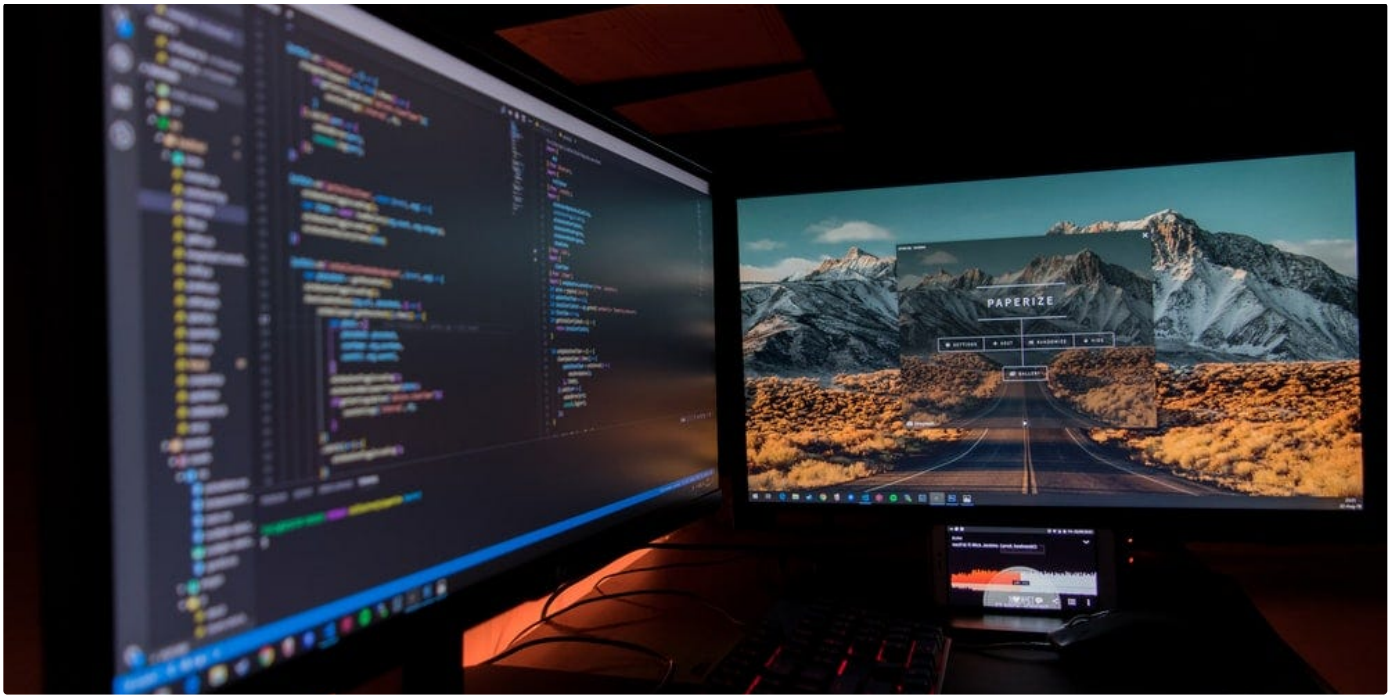
Follow

Written by Dr. Saptarsi Goswami

158 Followers · Writer for Towards Data Science

Advisor - AchieveX Solutions Pvt Ltd ,Asst Prof — CS Bangabasi Morning Clg, Lead Researcher University of Calcutta Data Science Lab, ODSC Kolkata Chapter Lead

More from Dr. Saptarsi Goswami and Towards Data Science



Dr. Saptarsi Goswami in Towards Data Science

How to use a pre-trained model (VGG) for image classification

why reinvent the wheel

5 min read · Oct 26, 2020

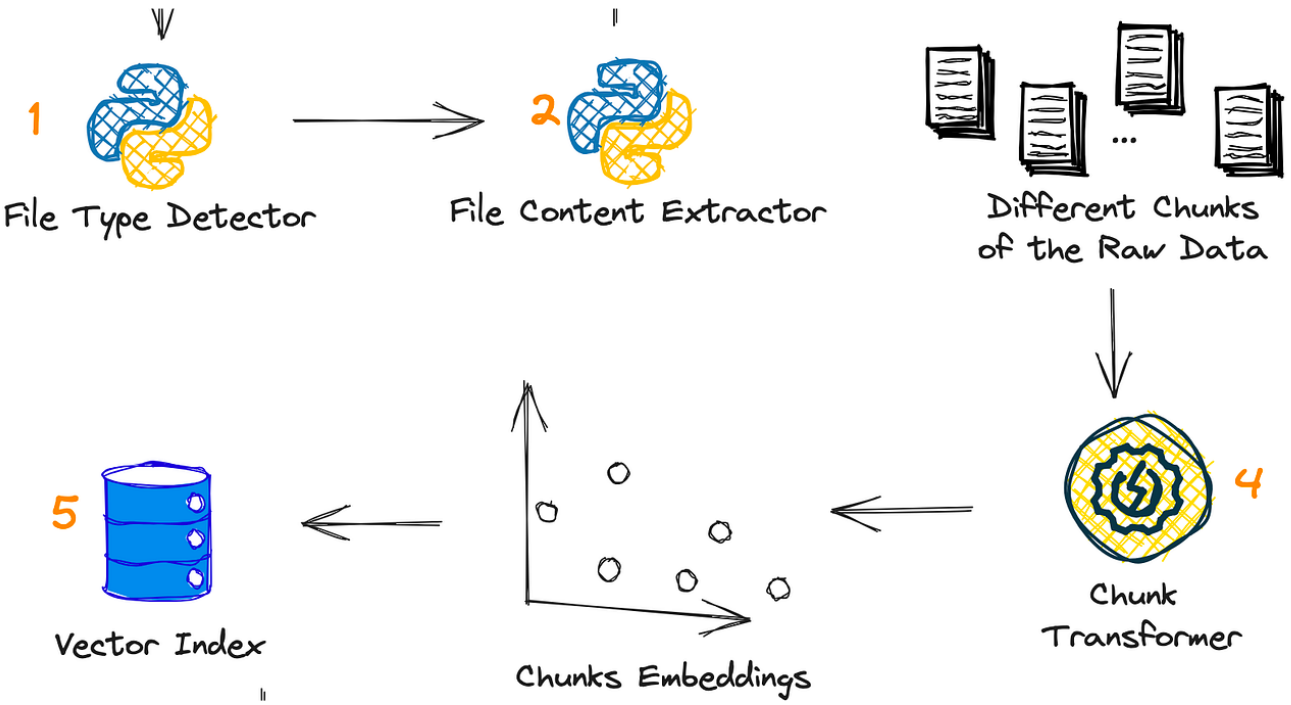
Bex T. in Towards Data Science

130 ML Tricks And Resources Curated Carefully From 3 Years (Plus Free eBook)

Each one is worth your time

✦ · 48 min read · Aug 1

3.3K 11



Zoumana Keita in Towards Data Science

How to Chat With Any File from PDFs to Images Using Large Language Models—With Code

Complete guide to building an AI assistant that can answer questions about any file

✦ · 9 min read · Aug 5

1.1K 12

Dr. Saptarsi Goswami in Towards Data Science

Using the Chi-Squared test for feature selection with implementation

The fewer the features, the easier to interpret the model

5 min read · Nov 13, 2020

AkShit Singh in Dev Genius

Image Classification on Imbalanced Dataset #Python #MNIST_dataSet

Image classification can be performed on an Imbalanced dataset, but it requires additional considerations when calculating performance...

8 min read · Apr 12

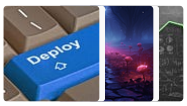
Barak Or, PhD in Towards Data Science

Solving The Class Imbalance Problem

Class imbalance is a common issue where the distribution of examples within a dataset is skewed or biased.

8 min read · Jan 4

Lists

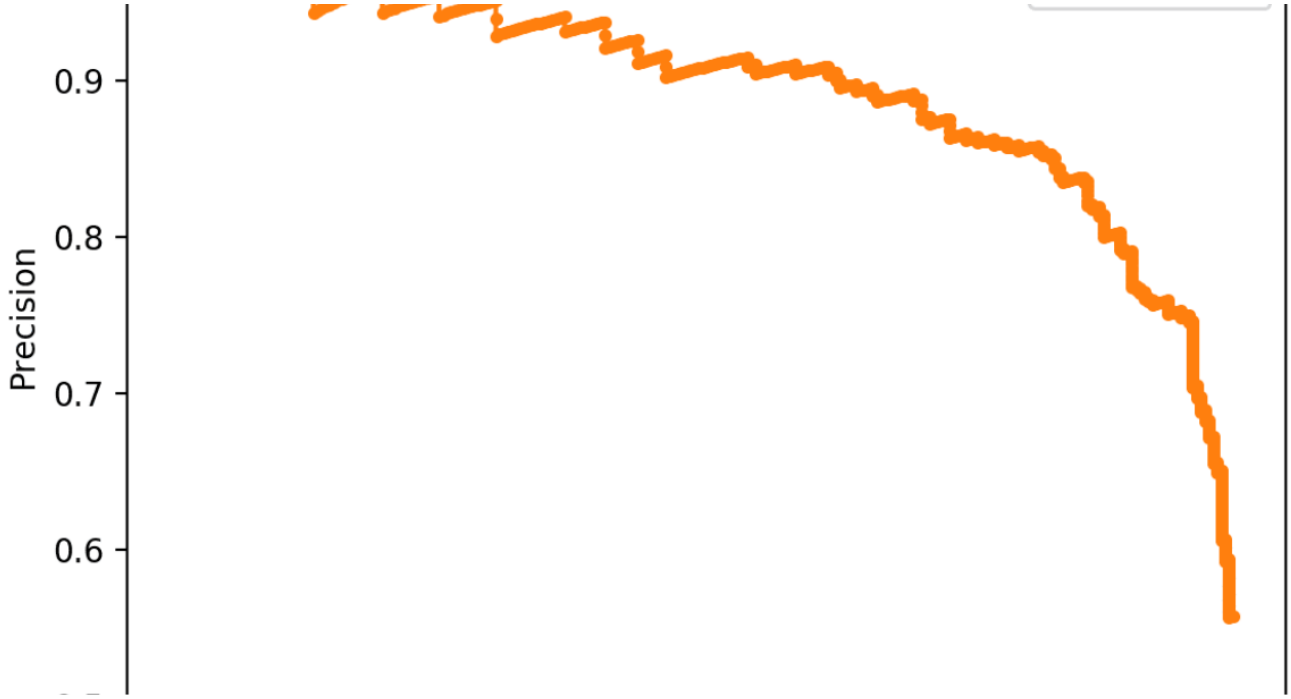


Predictive Modeling w/ Python
20 stories · 328 saves

New_Reading_List
174 stories · 89 saves

Practical Guides to Machine Learning
10 stories · 365 saves

Coding & Development
11 stories · 138 saves



Dr Mabrouka Abuhmida

Imbalanced datasets

by Dr Mabrouka Abuhmida

3 min read · Jul 10

Likhitha

All methods for balancing imbalanced data

Balancing data in machine learning is to address class imbalance, where one class has significantly fewer samples than another. Imbalanced...

2 min read · Aug 7

50

Aicha Bokbot in Artificial Intelligence in Plain English

4 ways to encode categorical features with high cardinality

We explore 4 methods to encode categorical variables with high cardinality: target encoding, count encoding, feature hashing and embedding.

· 9 min read · Jun 26

16

Arpita Kaushik

Oversampling on Imbalanced dataset

Imbalance Dataset

3 min read · May 24