ROC curve and precision-recall curves

1. Each plot can be summarized with an area under the curve score for model comparison
2. Trade-off in performance for different threshold values
3. ROC curve and ROC AUC can be optimistic on severely imbalanced classifications

Minority class is the positive outcome

Roc curve: plot of false positive rate vs true positive rate

*AUCROC can be interpreted as the probability that the scores given by a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.*

— Page 54, [Learning from Imbalanced Data Sets](https://amzn.to/307Xlva), 2018.

For oversampling, precision recall curve

“ROC analysis does not have any bias toward models that perform well on the majority class at the expense of the majority class – a property that is quite attractive when dealing with imbalanced data.

Imblanced learning: foundations, algorithms, and applications, 2013

*Although ROC graphs are widely used to evaluate classifiers under presence of class imbalance, it has a drawback: under class rarity, that is, when the problem of class imbalance is associated to the presence of a low sample size of minority instances, as the estimates can be unreliable.*

— Page 55, [Learning from Imbalanced Data Sets](https://amzn.to/307Xlva), 2018.

Although widely used, the ROC AUC is not without problems.

For imbalanced classification with a severe skew and few examples of the minority class, the ROC AUC can be misleading. This is because a small number of correct or incorrect predictions can result in a large change in the ROC Curve or ROC AUC score.

*#在某些情况 下， ROC 曲线 不受 类别个数 影响 的 优点也是其缺点。*

*#正如 上述的例子中，当 负 样本增加了 10 倍 之后，从 ROC 曲线 上无法 反映 出 变化 。*

*# 从下图的混淆 矩阵中， 我们可以 明显的看出两种情况下的区别。 可以 看到此时 实际为负，#预测为正的样本 点大量增加（ 即 FP ） ， 这将导致 正 类的 精准率（ Precision ） 显着 的降低 。# 从分类 报告中可以看到 正 类的 Precision 从 0.94 降低 到了 0.14 。*

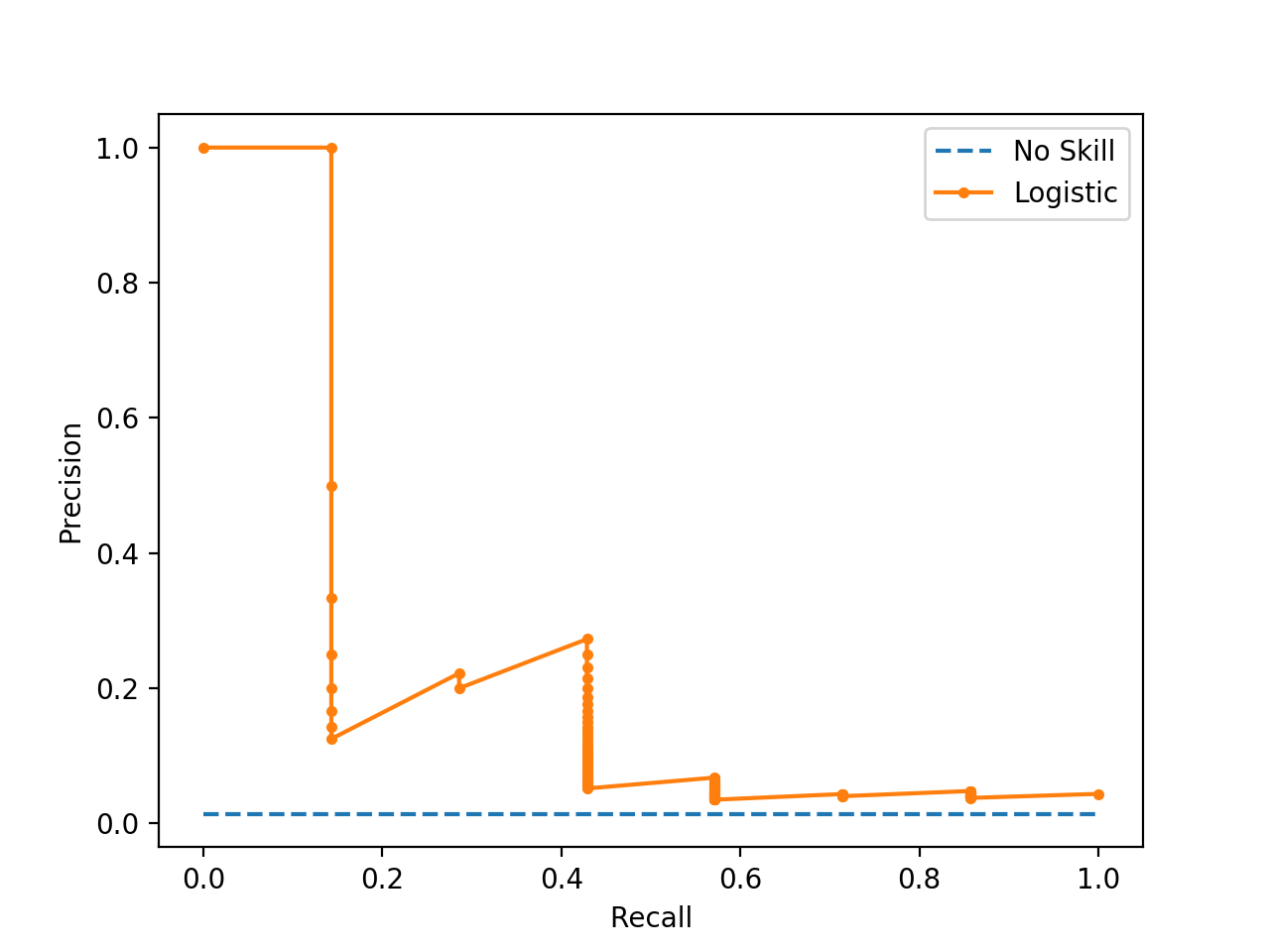
*# 也就是 说，在 极 不平衡 的 数据集下， ROC 曲线会给出 一个 过于 乐观的 估计*

*Although ROC graphs are widely used to evaluate classifiers under presence of class imbalance, it has a drawback: under class rarity, that is, when the problem of class imbalance is associated to the presence of a low sample size of minority instances, as the estimates can be unreliable.*

— Page 55, [Learning from Imbalanced Data Sets](https://amzn.to/307Xlva), 2018.

A common alternative is the precision-recall curve and area under curve.

A model with perfect skill is depicted as a point at a coordinate of (1,1). A skillful model is represented by a curve that bows towards a coordinate of (1,1). A no-skill classifier will be a horizontal line on the plot with a precision that is proportional to the number of positive examples in the dataset. For a balanced dataset this will be 0.5.



We can see the horizontal line of the no skill classifier as expected and in this case the zig-zag line of the logistic regression curve close to the no skill line.

To explain why the ROC and PR curves tell a different story, recall that the PR curve focuses on the minority class, whereas the ROC curve covers both classes

He, H., & Ma, Y. (Eds.). (2013). *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

1. Using the weights parameters in Sci-Kit Learn classifiers

Over and Undersampling

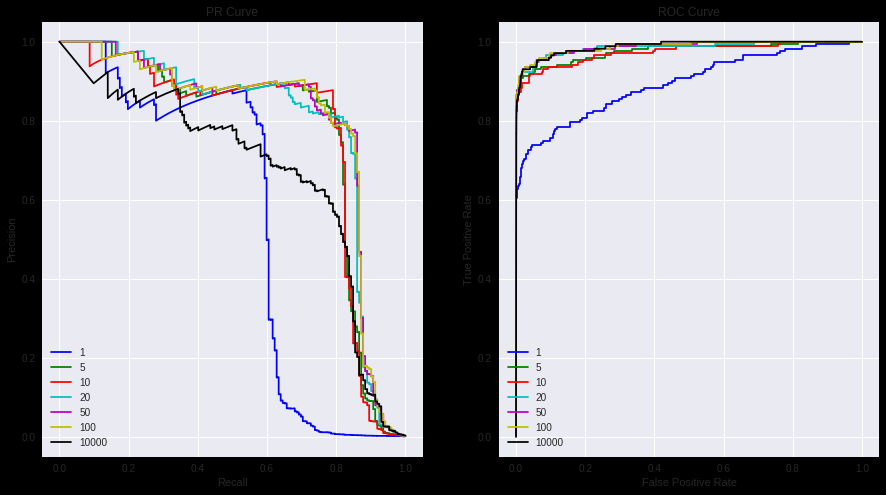
class using a simple parameter during model initiation. Let's see how that will improve our results

In [14]:

lr\_model = LogisticRegression(class\_weight='balanced')

have increased our Fraud recall score at the expense of more mis-classified Legit cases. With the "balanced" weight parameter, we have increased our false positive counts from 39 to 2300. 2300 is still only a small fraction of truely negative cases (out of 99511),

we have increased our Fraud recall score at the expense of more mis-classified Legit cases.



While the blue, w=1, line performed poorly in both charts, the black, w=10000, line performed "well" in the ROC but poorly in the PR curve. This is due to the high class imbalance in our data. ROC curve is not a good visual illustration for highly imbalanced data, because the False Positive Rate ( False Positives / Total Real Negatives ) does not drop drastically when the Total Real Negatives is huge Whereas Precision ( True Positives / (True Positives + False Positives) ) is highly sensitive to False Positives and is not impacted by a large total real negative denominator.

The biggest difference among the models are at around 0.8 recall rate. Seems like a lower weight, i.e. 5 and 10, out performs other weights significantly at 0.8 recall. This means that with those specific weights, our model can detect frauds fairly well (catching 80% of fraud) while not annoying a bunch of customers with false positives with an equally high precision of 80%.

Recall = true positive rate

ROC: balanced on positive and negative instance (TPR focus on positive cases, and FPR focus on negative case)

Good: TPR only rely on P (positive cases) and FPR reply on only Negative cases， so when the negative cases are 10 times more, it doe not affect the ROC

BAD: negative cases increased a lot but curve is not changing, which means a lot of false positives are generated (treat negative as positive )

FPR = FP / N = FP/(FP+TN) . When N >> P , FP increase only result in small changes in FRP. But it should affect precision-recall curve a lot

Precision- Recall curve

Also used Recall but precision focused on the positive case

相同类别分布下 正例的估计 应当使用

在曲线上找到最优的点 得到相对于的preciion recall f1 score ,调整阈值， 得到符合具体应用 的模型

Davis and Goadrich propose PR ove ROC

Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In  
ICML’06: Proc. of the 23rd Int. Conf. on Machine Learning (ACM ICPS). ACM, New York, NY, 233–  
240.

*Precision-recall curves* (*PR curves*) are recommended for highly skewed domains  
where *ROC* curves may provide an excessively optimistic view of the performance [Davis and Goadrich 2006]

Better to highlight differences between models for highly imbalanced data

i.e. PR more sensitive

example dependent Cost-sensitive learning

cost-sensitive training vs cost dependent classification

1. Train a model with a loss function that minimizes the actual costs, instead of the misclassification errors （cost-sensitive training）
2. Train a regular model, but classify each sample according to the lowest expected costs (cost classification models)

Cost\_fp = 3

Cost\_fn = data[‘amout’]

Cost\_tp = 3

Cost\_tn = 3

不过在这里我不得不得出一个比较悲观的结论：就这两个数据集的结果来看，如果本身数据偏斜不是很厉害，那么采样方法的提升效果很细微。如果本身数据偏斜很厉害，采样方法纵使比base model好很多，但由于base model本身的少数类预测能力很差，所以本质上也不尽如人意。这就像考试原来一直靠10分，采样了之后考了30分，绝对意义上提升很大，但其实还是差得远了。

The SMOTE algorithm has been applied with several different classifiers and was  
also integrated with boosting

they are insensitive to changes in class distribution. If the proportion of  
positive to negative instances changes in a test set, the ROC curves will not change

classweights

1:1000