北京大学

企业破产预测模型：设计与实现

机器学习概论课程项目期末 报告

张柏舟 周裕涵 宋铭宇 柯佳奇 鲁琦琨

内容提要

- 数据降维
- 平衡学习
- 贝叶斯网学习
- 结果对比分析

数据降维

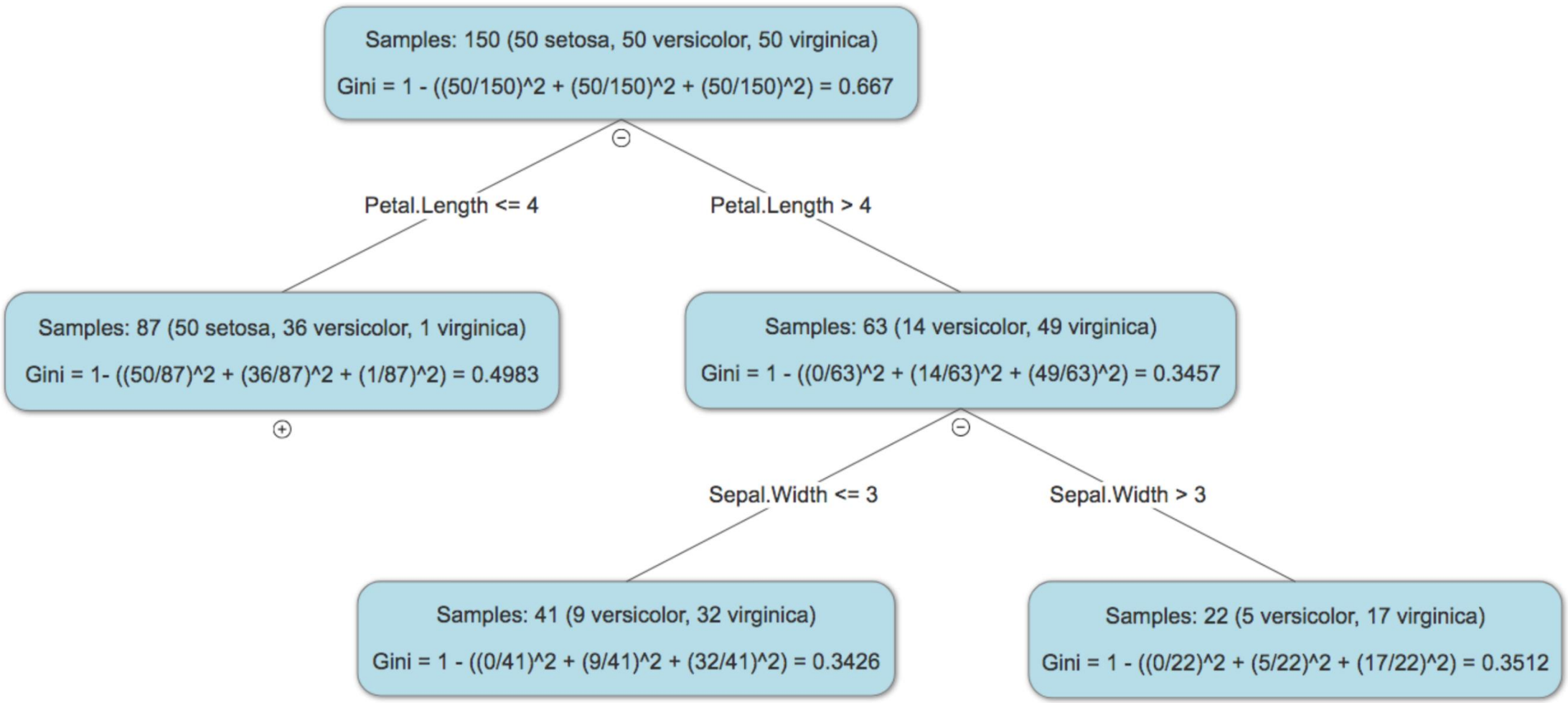
- Gini Importance
- Principal Component Analysis (PCA)
- Economical Analysis

数据降维：Gini Importance

- Gini Importance: 也称Mean decrease in impurity, 常用于含有决策树的模型，例如随机森林
- Gini Importance是Sklearn内置的feature_importance函数的实现方法。
- 概念复习：用gini_index来衡量数据集D的purity

$$\begin{aligned}\mathrm{Gini}(D) &= \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'} \\ &= 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2 \ .\end{aligned}$$

数据降维：Gini Importance

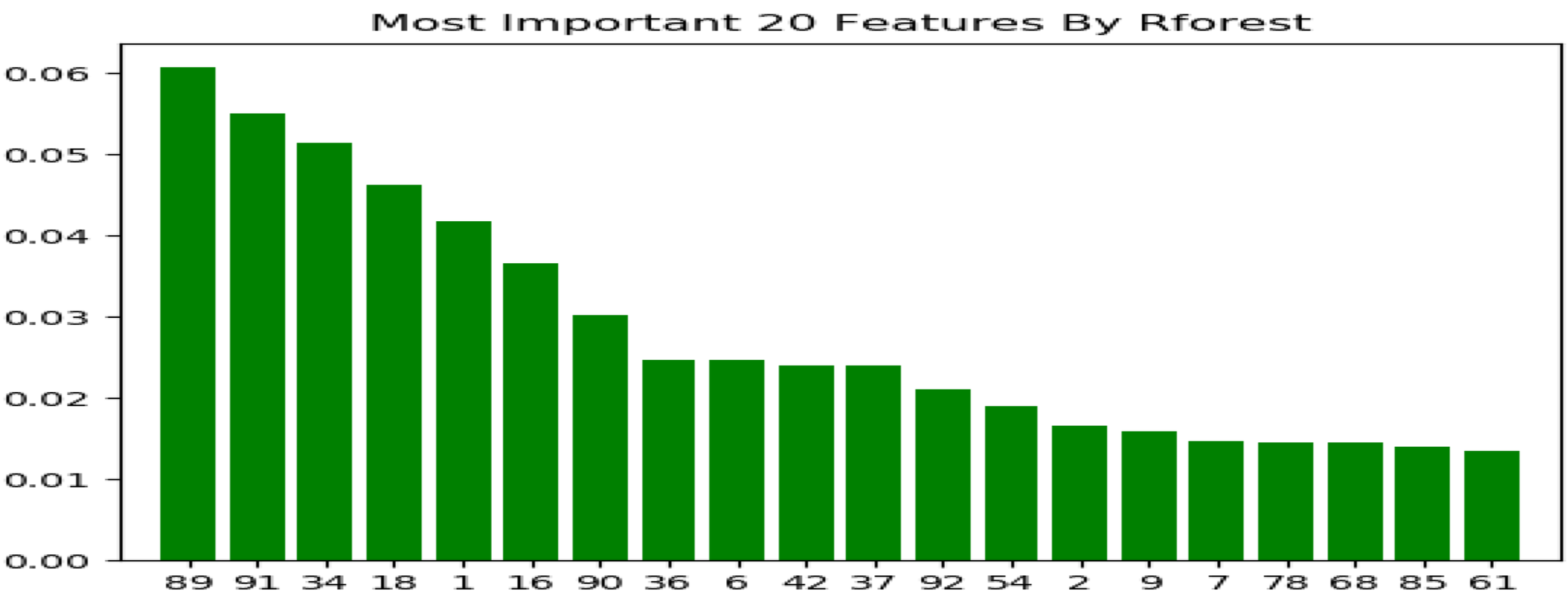


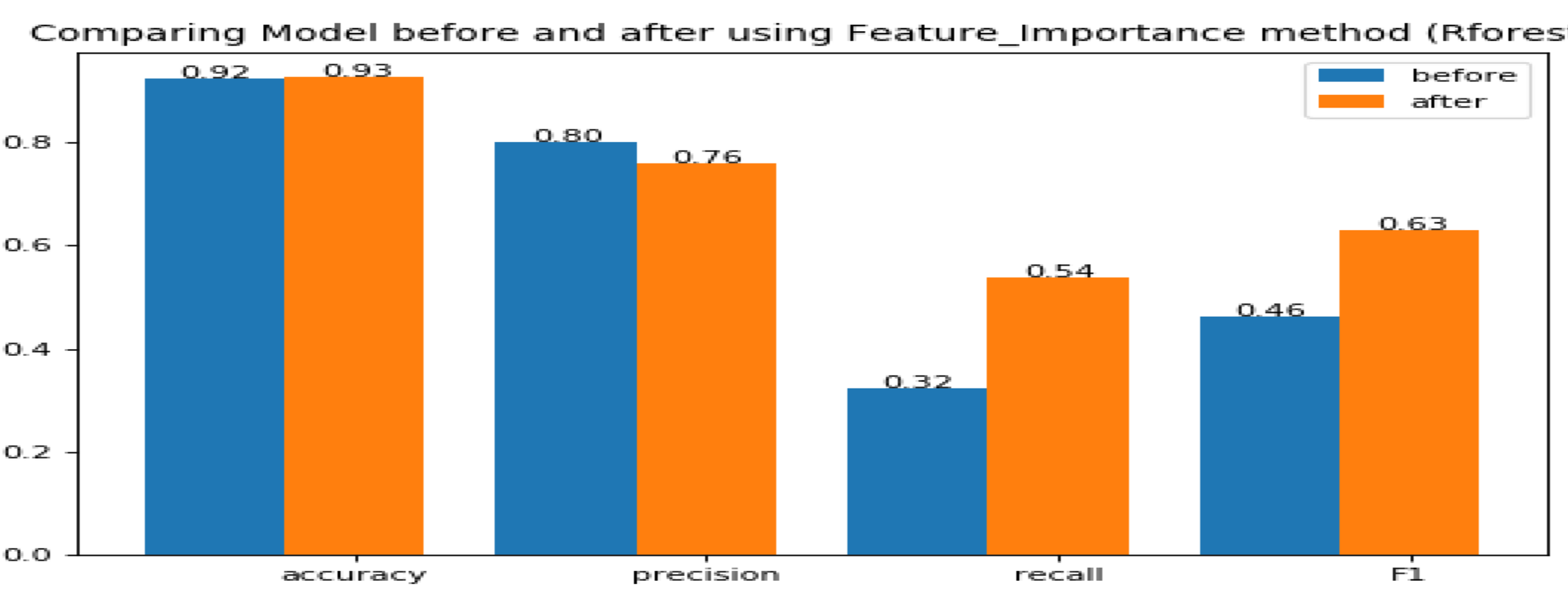
- 在某一个节点，决策树依据被分为两棵子树，Decrease in impurity=Gini(parent) -weighted_sum(Gini(childs))
- Gini Importance: 一个属性在所有可能的决策树中的decrease in impurity 取均值

数据降维：Gini Importance

- 使用多种含决策树的预测模型（Decision Tree、Random Forest、Gradient Boost、等）进行特征工程，以Gini Importance的大小为标准（越大越好），得到前20个最重要特征。

- 不同模型筛选出的特征大部分相同，且采用这些特征重新训练后，预测准确率、召回率、F1指数都有不同程度增减


Most Important 20 Features By Rforest
0.06
0.05
0.04
0.03
0.02
0.01
0.00
89 91 34 18 1 16 90 36 6 42 37 92 54 2 9 7 78 68 85 61


Comparing Model before and after using Feature_Importance method (Rforest)
before
after
0.92
0.93
0.80
0.76
0.32
0.54
0.46
0.63
0.8
0.6
0.4
0.2
0.0
accuracy
precision
recall
F1

数据降维：PCA

- PCA: 主成分分析法

Principal Component Analysis (PCA) algorithm summary

→ After mean normalization (ensure every feature has zero mean) and optionally feature scaling:

$$\text{Sigma} = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)})(x^{(i)})^T$$

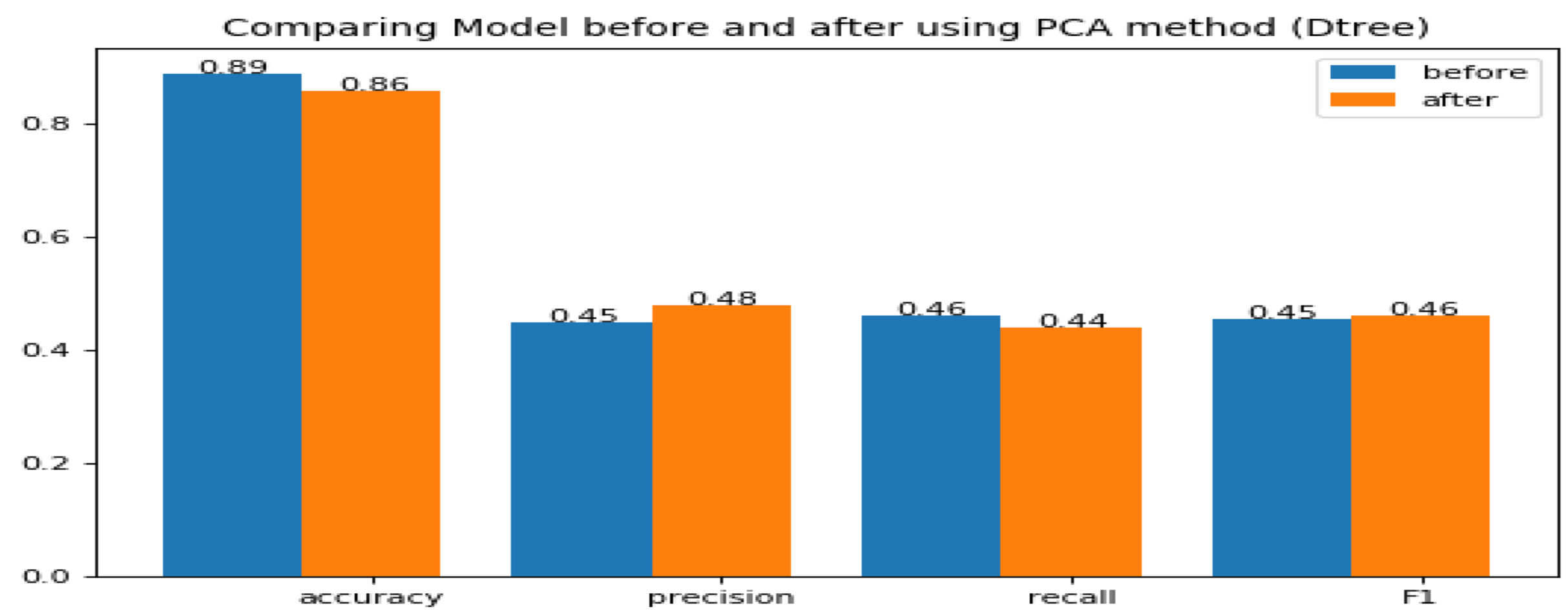
```
→ [U,S,V] = svd(Sigma);
→ Ureduce = U(:,1:k);
→ z = Ureduce'*x;
```

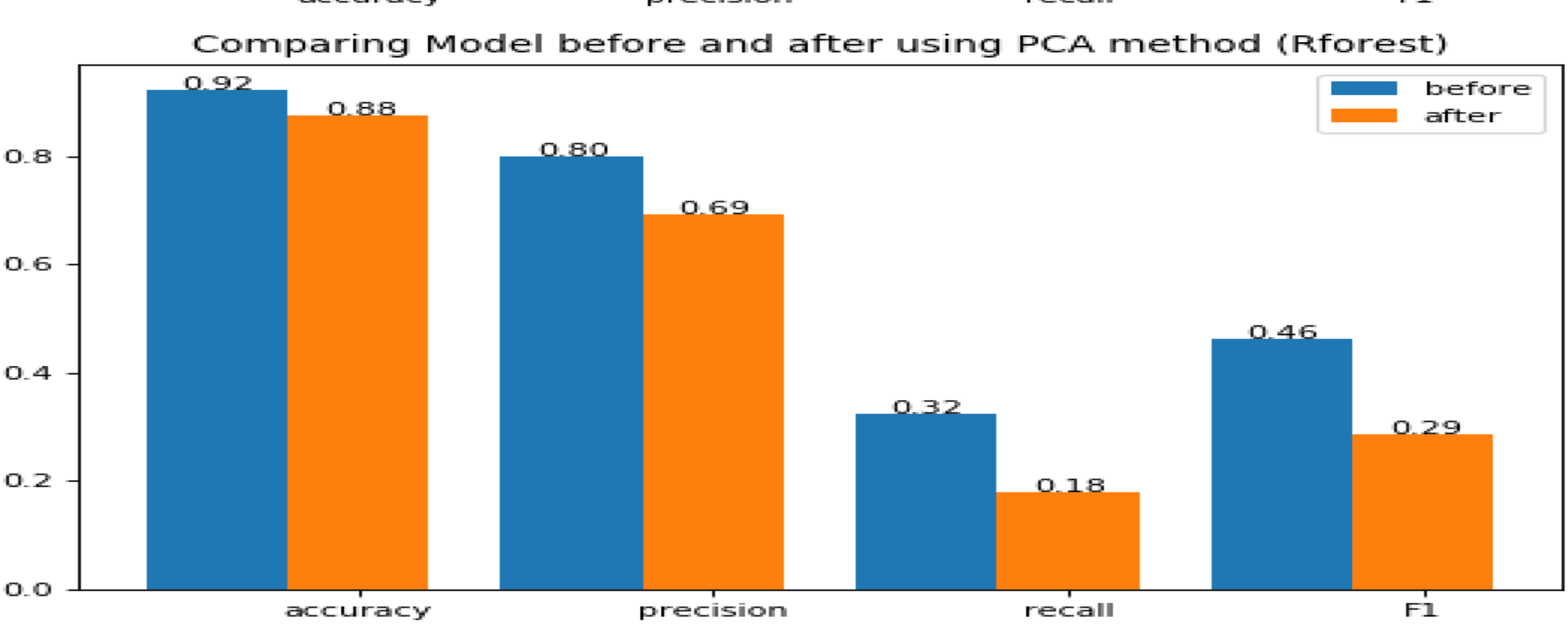
$X = \begin{bmatrix} — & x^{(1)T} & — \\ & \vdots & \\ — & x^{(m)T} & — \end{bmatrix}$

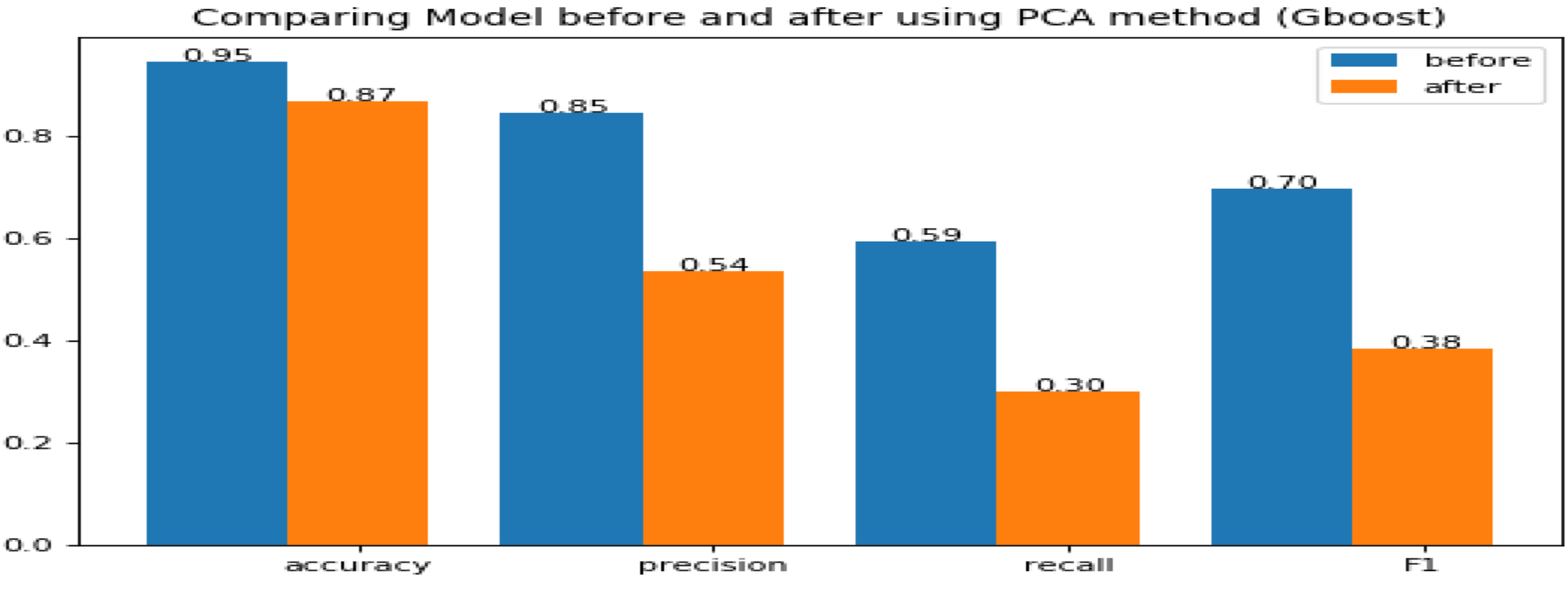
Sigma = (1/m) * X' * X;

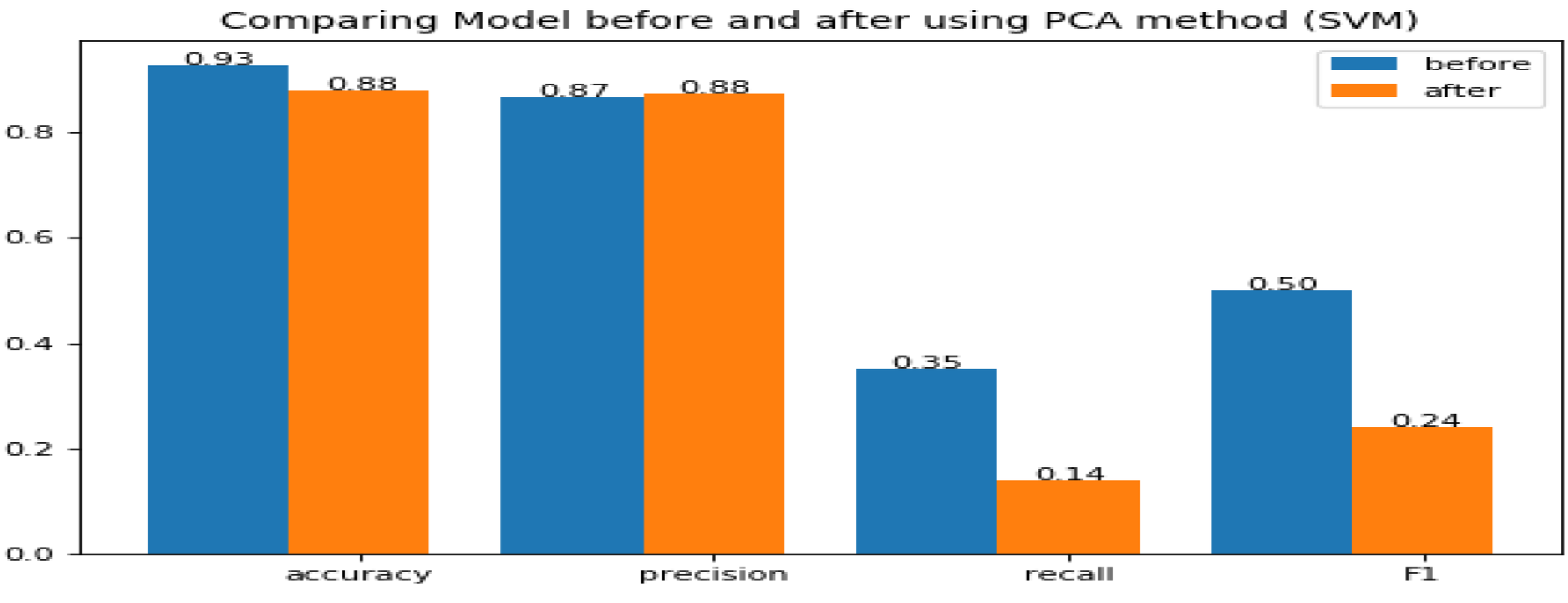
$x \in \mathbb{R}^n$ $x_0 = 1$

图片来源：AndrewNg Machine Learning 课件


Comparing Model before and after using PCA method (Dtree)
before
after
0.89
0.86
0.45
0.48
0.46
0.44
0.45
0.46
0.0
0.2
0.4
0.6
0.8
accuracy
precision
recall
F1


Comparing Model before and after using PCA method (Rforest)
before
after
0.92
0.88
0.80
0.69
0.32
0.18
0.46
0.29
0.0
0.2
0.4
0.6
0.8
accuracy
precision
recall
F1


Comparing Model before and after using PCA method (Gboost)
before
after
0.95
0.87
0.85
0.54
0.59
0.30
0.70
0.38
0.0
0.2
0.4
0.6
0.8
accuracy
precision
recall
F1


Comparing Model before and after using PCA method (SVM)
before
after
0.93
0.88
0.87
0.88
0.35
0.14
0.50
0.24
0.0
0.2
0.4
0.6
0.8
accuracy
precision
recall
F1

数据降维：PCA

- PCA为何导致四个指标普遍下降:

PCA 原理主要是之间的相关性，并且假设这种相关性是线性的，对于非线性的依赖关系则不能得到很好的结果。使用PCA容易导致非线性依赖关系的丢失，从而降低模型性能。

- PCA 方法不采用

数据降维：Economical Aalysis

1. ROA(A) before interest and % after tax

净利润与总资产的比值，描述公司赚钱的能力，即公司的每一块钱能赚多少钱

3. Operating Gross Margin　5. Operating Profit Rate

描述公司经营性业务的盈利能力，3为毛利率、5为净利率

12. Cash Flow Rate

现金流，长期运营下去的重要指标

13. Interest-bearing debt interest rate

负息债务的利息,利息越高，公司越有可能破产

16. Net Value Per Share (A)

(总资产-总负债) / 股数，大概描述了公司的负债情况

18. Persistent EPS in the Last Four Seasons

过去一年里的每股净收益,衡量公司盈利能力

31. Cash Reinvestment

现金再投资比率，企业能用于再投资的现金是多还是少?

32. Current Ratio　33. Quick Ratio

流动比率（流动资产 / 流动负债），公司会不会陷入短期的流动性问题

35. Total Debt　36. Debt ratio %　94. Equity to Liability

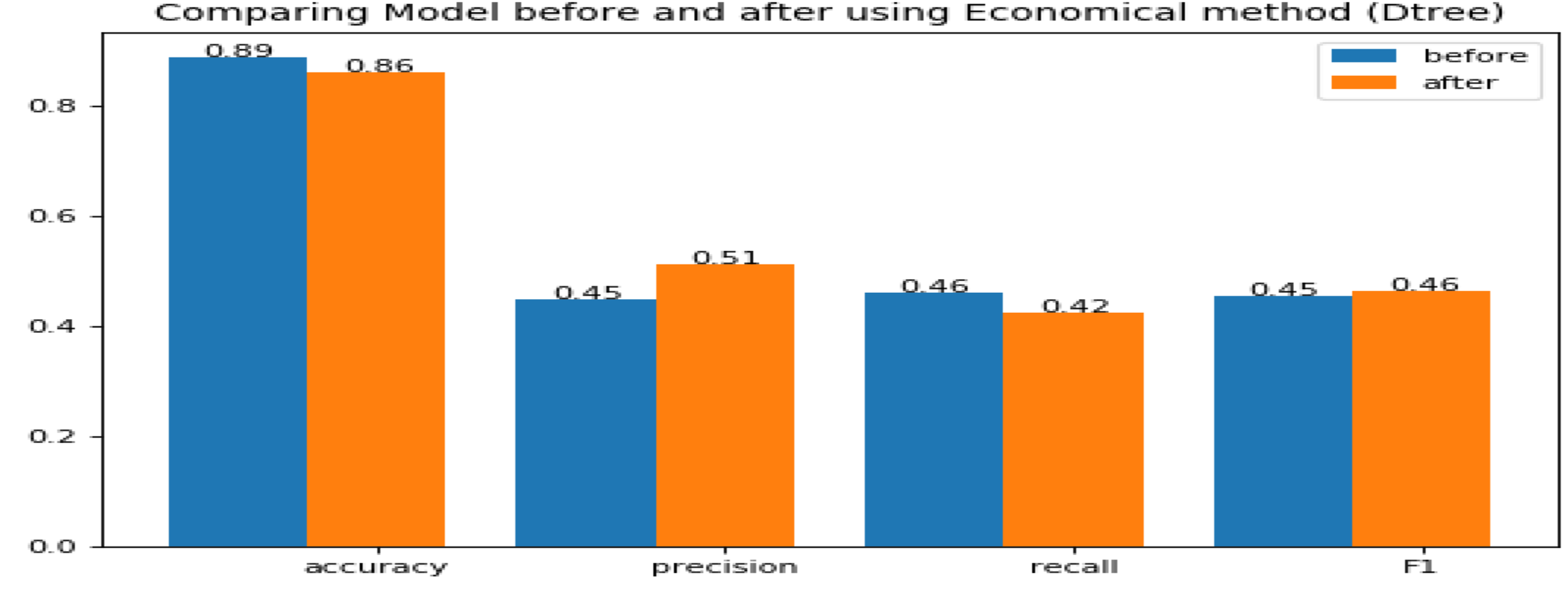
衡量借债的比率

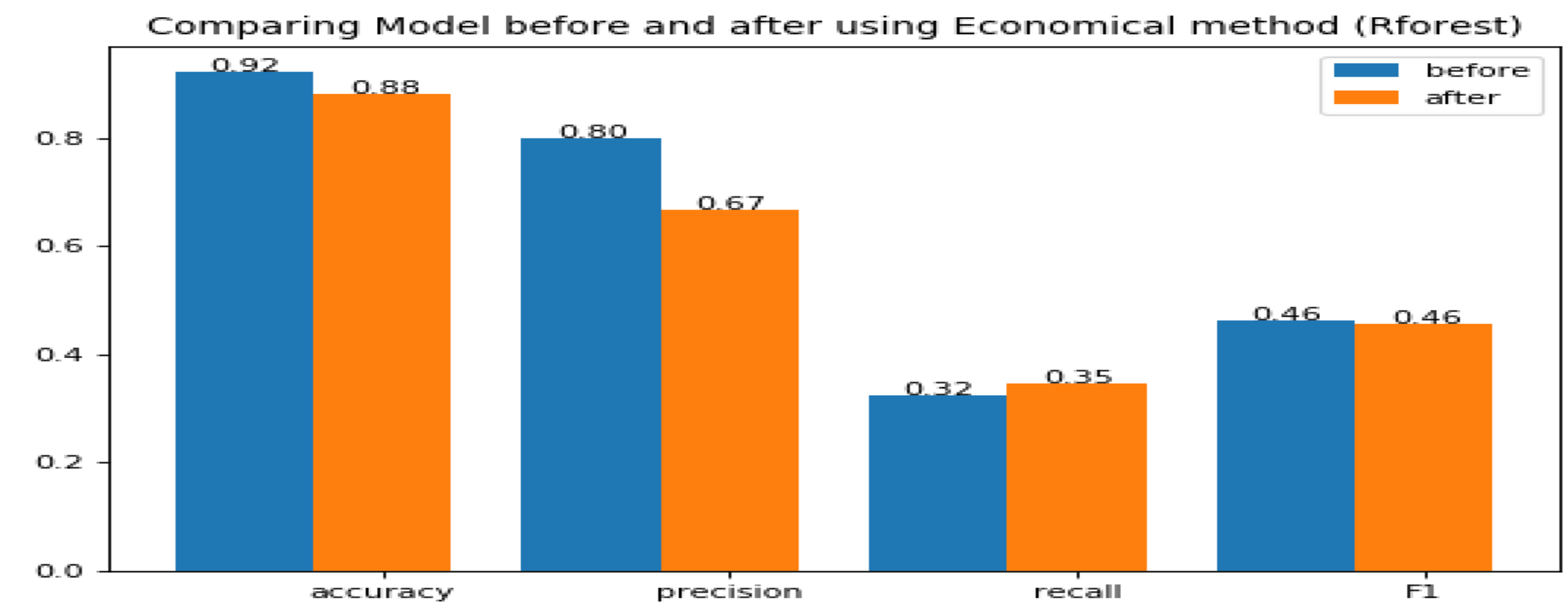
41. Operating profit　42. Net profit before tax

股东提供的每单位资金产生多少利润和税前净利润

92. ebit: earning before interest and tax

赚的钱有多少用来付利息和税务，如果值太高意味着公司可能会破产


Comparing Model before and after using Economical method (Dtree)
before
after
0.89
0.86
0.45
0.51
0.46
0.42
0.45
0.46
0.0
0.2
0.4
0.6
0.8
accuracy
precision
recall
F1


Comparing Model before and after using Economical method (Rforest)
before
after
0.92
0.88
0.80
0.67
0.32
0.35
0.46
0.46
0.0
0.2
0.4
0.6
0.8
accuracy
precision
recall
F1

数据降维：综合

- 从性能角度来看，选取性能最优的Gini importance方法。
- 从可解释性的角度来看，经济学法也是可以被接受的。
- 结合以上两种方法的结果，最终选定了被提取的特征
- 当前的数据集的规模不够大（6000多条数据），要想得到更好的模型和更高的f值，必须收集更多的数据。

平衡学习

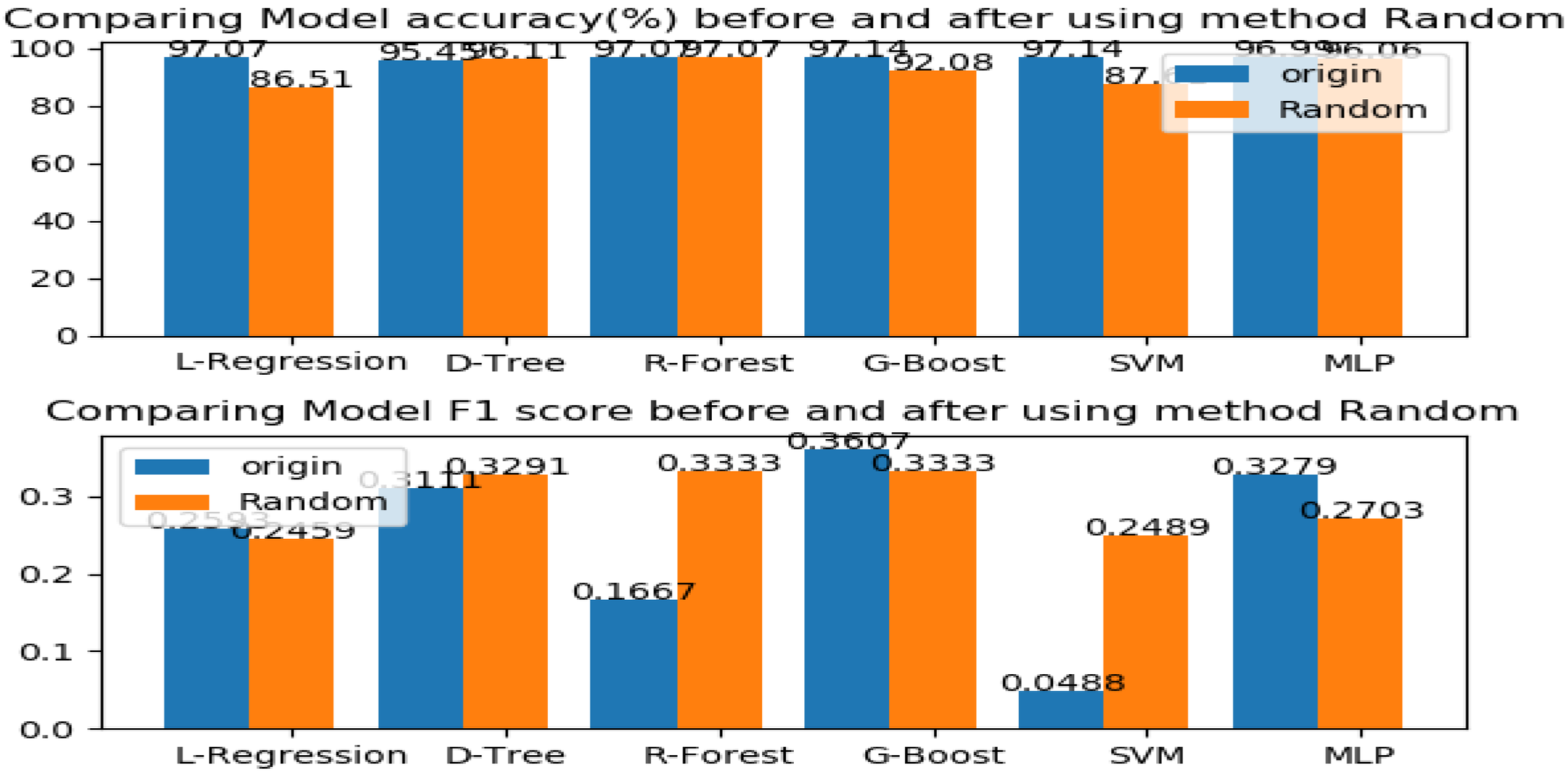
- 能够得到的破产公司数据很少，因此从真实世界得到的数据集中两类标签比例不均衡。使用平衡学习（Imbalanced Learning）处理

- 基本思想：

  - 利用已知数据构造合理的破产样本(over-sampling)
  - 减少非破产样本个数(under-sampling)
  - 前面两种结合

尝试一：随机过程采样（Random）

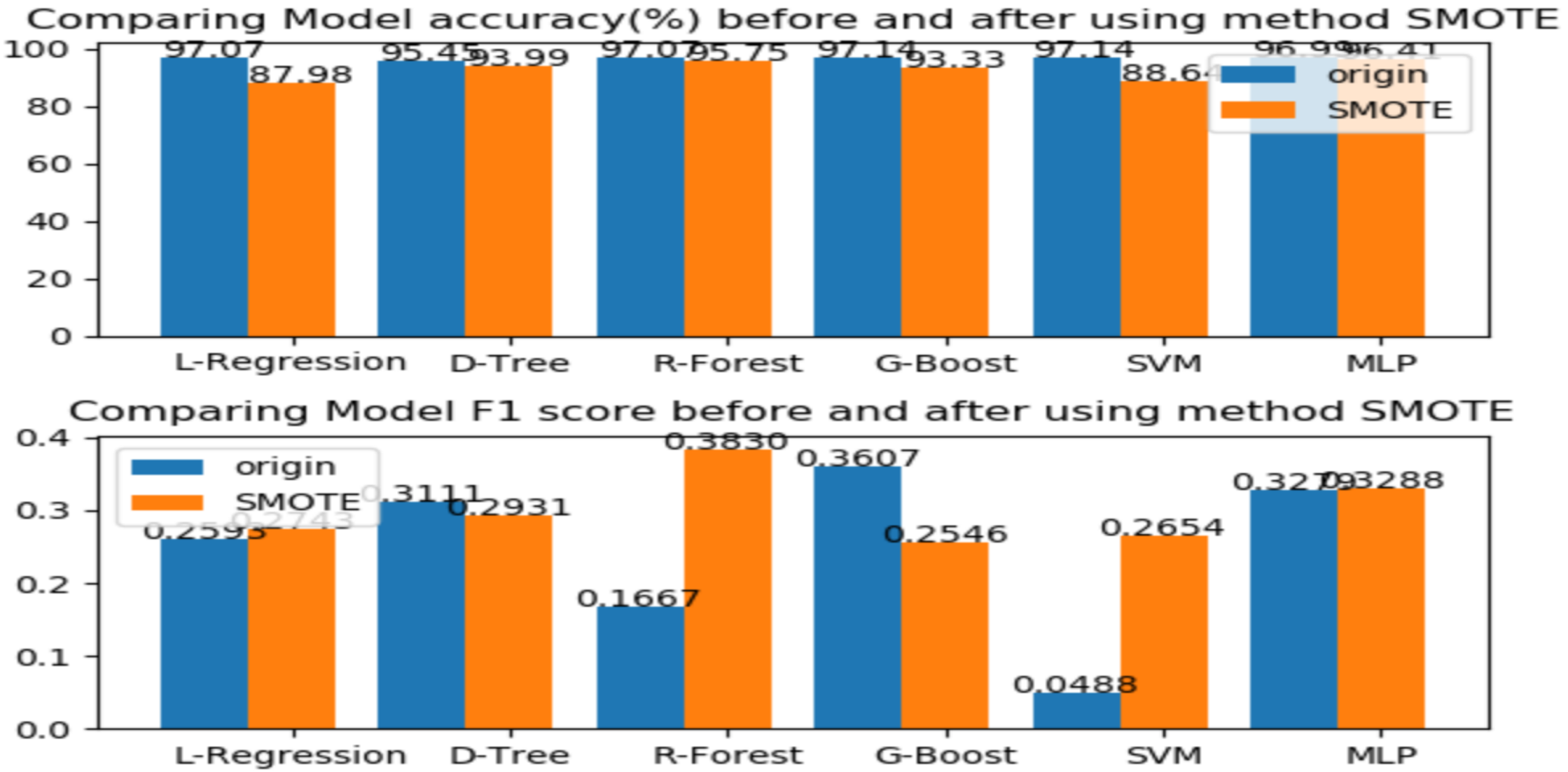
从少数类的样本中进行随机采样来增加新的样本，使得所有类样本数相同。



效果较差。可能原因：这个随机采样等价于给破产的样本赋予了很高的权重进行分类，导致这些破产的样本对模型影响太大，进而产生误差的积累。

尝试二：SMOT

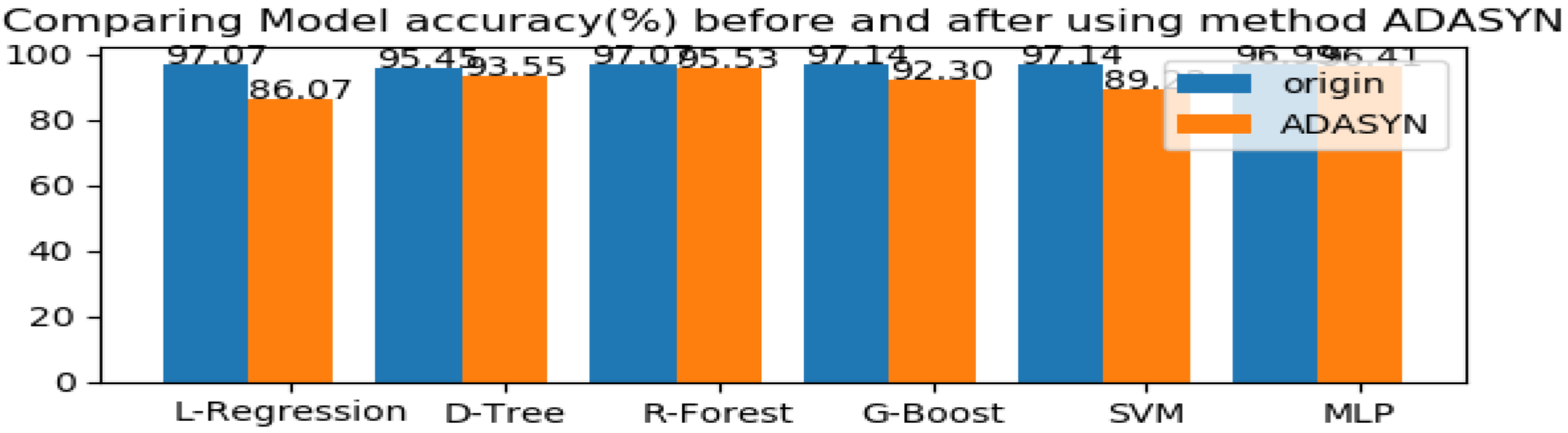
Synthetic Minority Oversampling Technique: 通过插值产生新的样本。对于少数类样本a, 随机选择一个近邻的少数类样本b, 然后对每一对a与b，从其连线上随机选取一个点c作为新的少数类样本，对每个样本a重复N次。

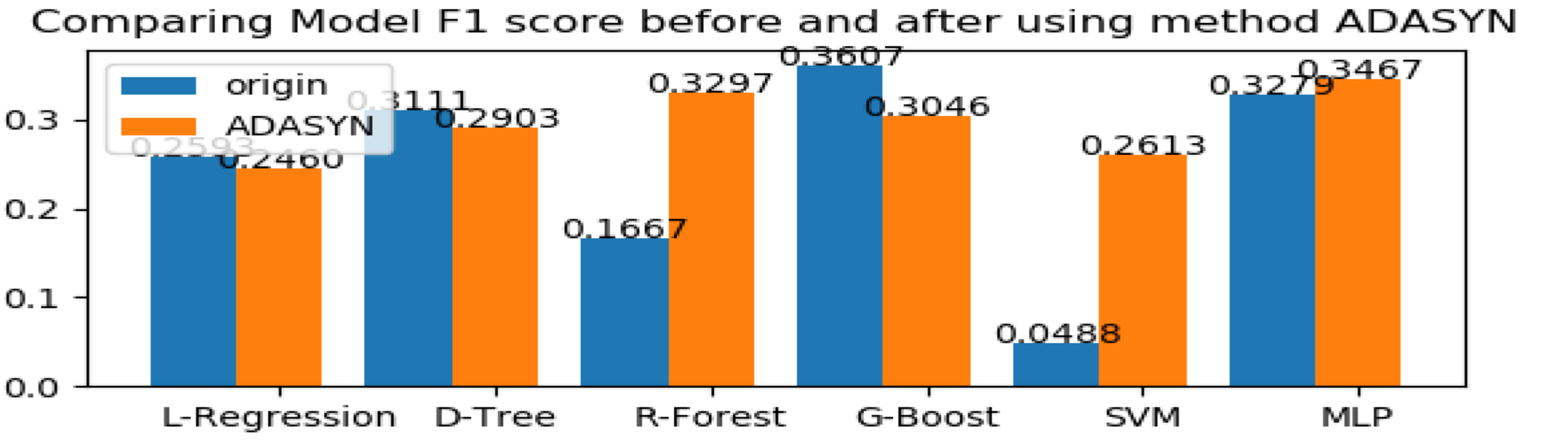


SMOTE对少数样本一视同仁，未考虑近邻样本的类别信息，往往出现样本混叠现象，导致分类效果不佳

尝试三：ADASYN

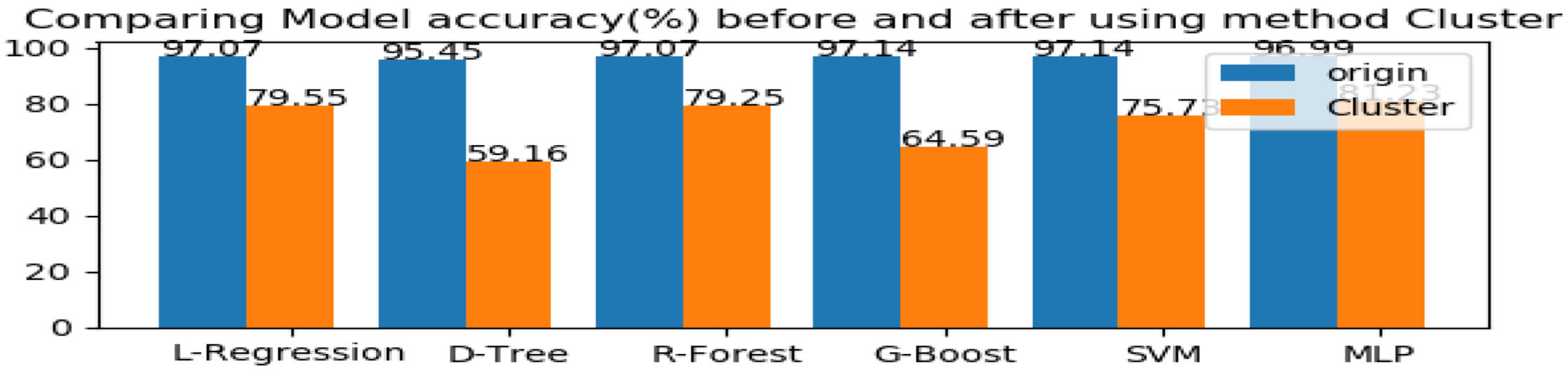
通过插值产生新的样本，与SMOTE中插值方法类似，但不同少数类样本根据其周围的多数类样本数目而进行不同数目的插值过程。K近邻中多数类样本数目越多，插值的次数越多。

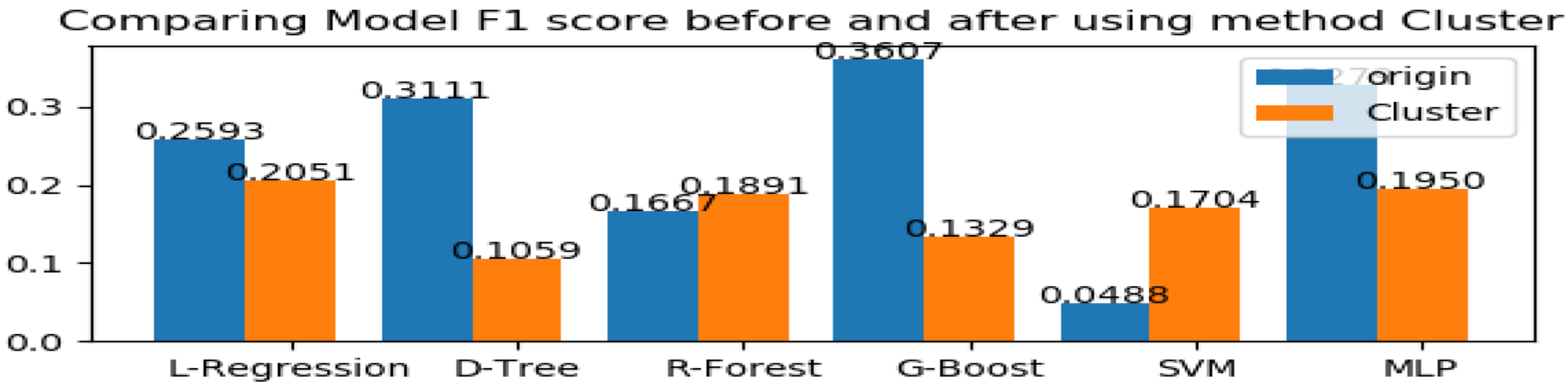




尝试四：ClusterCentroids

从原始数据生成若干数据，每一个类别（破产、不破产）的样本都会用K-means算法的中心点来进行合成，而不是随机从原始样本进行抽取

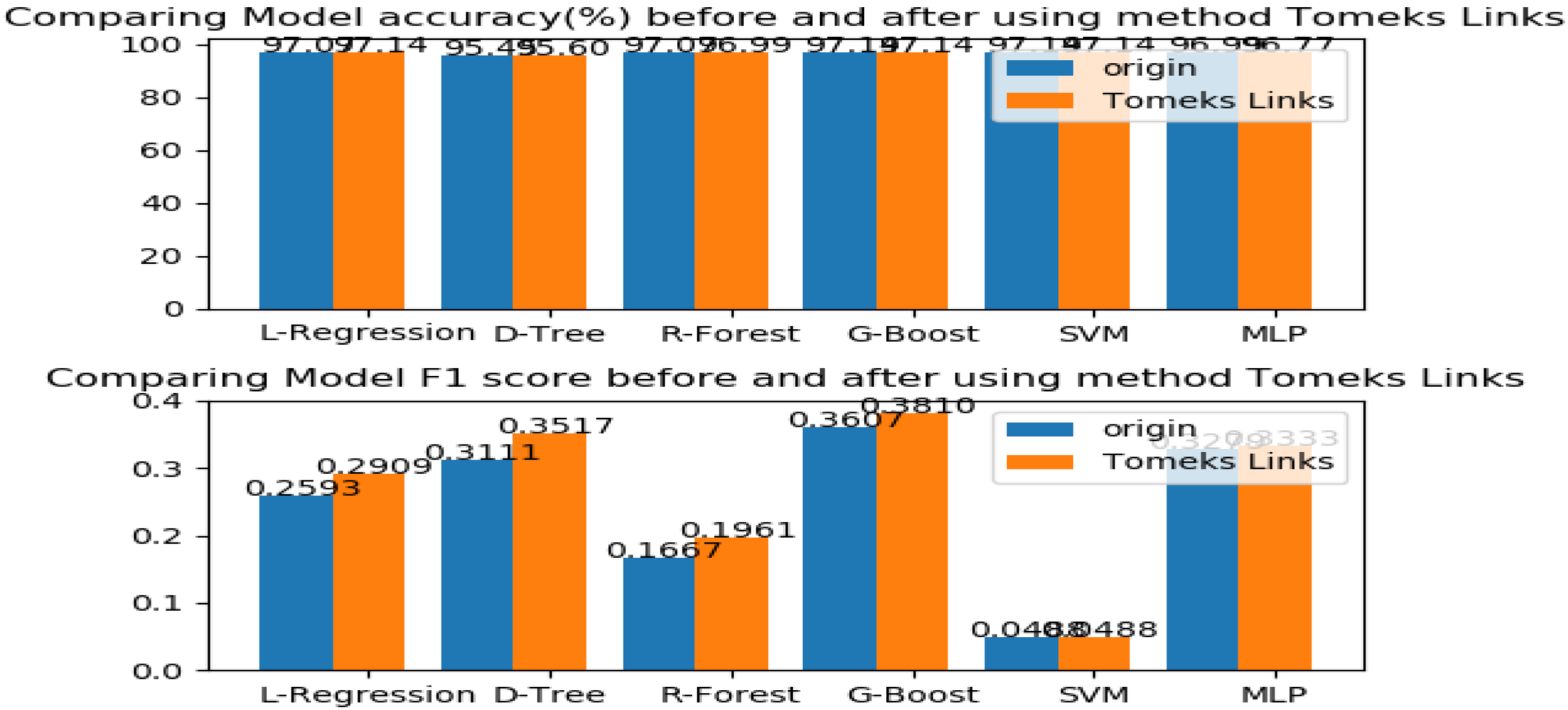




该方法要求原始数据集最好能聚类成簇，然而很明显这个数据并不适合聚类

尝试五：Tomek Links

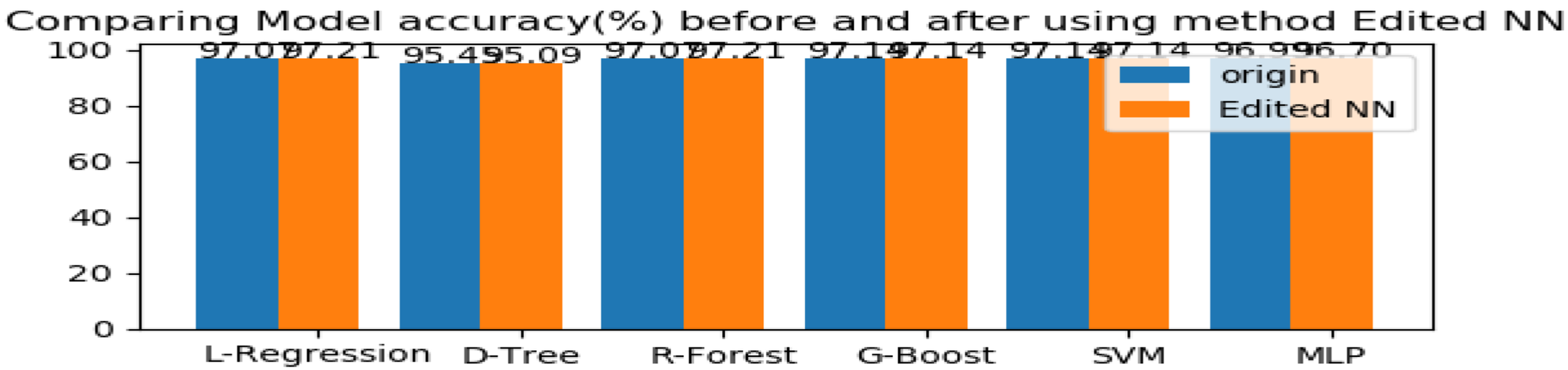
样本x与样本y来自于不同的类别, 满足以下条件, 它们之间被称之为Tomek Links：不存在另外一个样本z, 使得d(x,z) < d(x,y) 或者 d(y,z) < d(x,y)成立。这个时候, 样本x或样本y很有可能是噪声数据, 或者两个样本在边界的位置附近

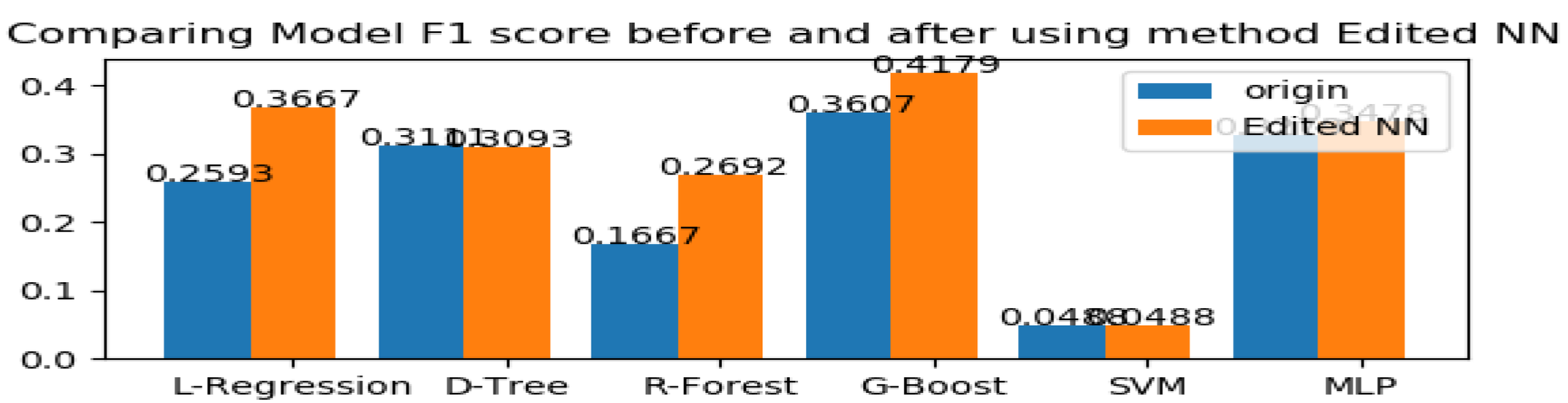


删除的样本数有限，得到结果和未经处理的情况几乎相同

尝试六：Edited Nearest Neighbours

对于属于多数类的一个样本，如果其K个近邻点有超过一半都不属于多数类，则这个样本会被剔除。这个方法的另一个变种是所有的K个近邻点都不属于多数类，则这个样本会被剔除

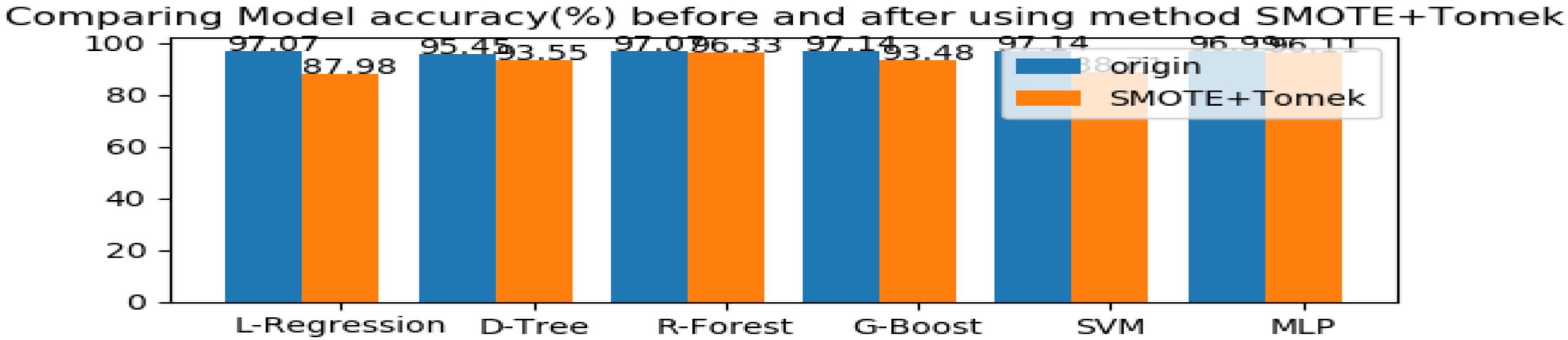


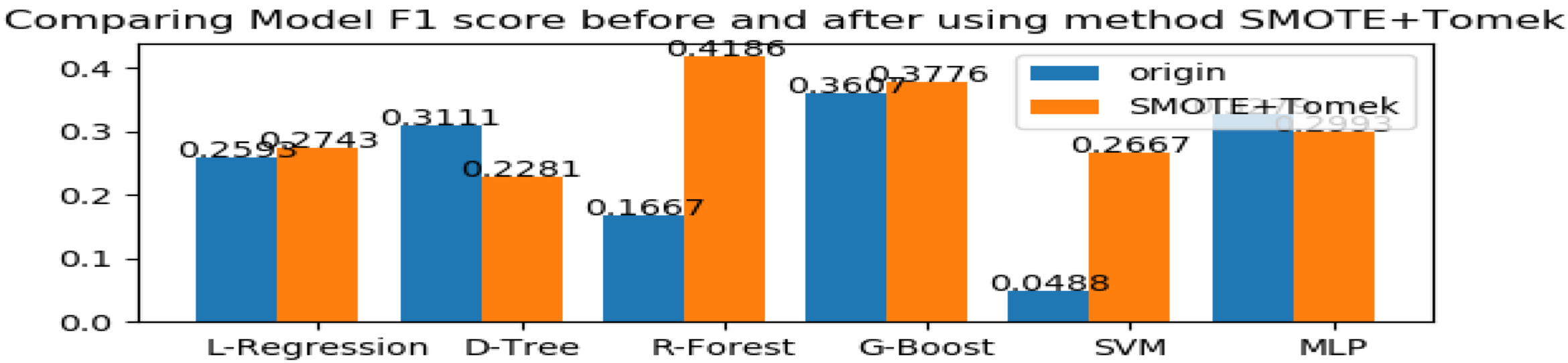


准确率较好，F1 score较未处理也有了一定的提升。

尝试七：SMOTE + Tomek Links

在之前的SMOTE方法中, 当由边界的样本与其他样本进行过采样差值时, 很容易生成一些噪音数据。因此, 在过采样之后需要对样本进行清洗。 而Tomek Links 与 Edited Nearest Neighbours方法都能实现上述的要求，这里只展示效果更好的 SMOTE + Tomek Links





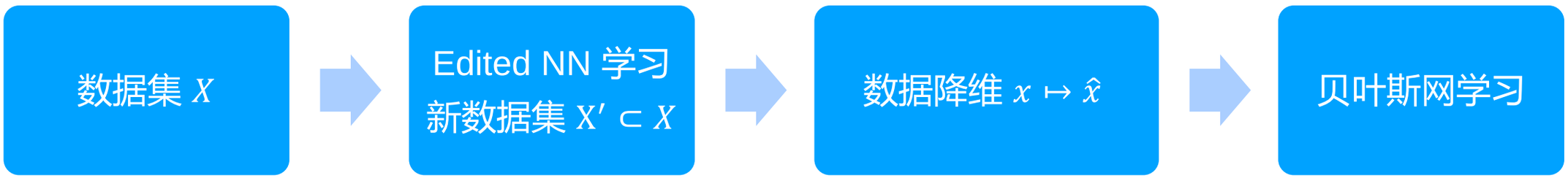
相比纯粹的SMOTE和Tomek Links效果有所改进

平衡学习：结论

- 实际上我们尝试了更多方法（如BorderlineSMOTE、更多over-sampling和undersampling的组合）
- 最终选定几乎维持了原准确率、对F1 score提升最好的Edited Nearest Neighbours 方法

贝特斯网络学习

- 输入特征之间存在依赖关系，朴素贝叶斯忽略了这种关系，而贝叶斯网络可以刻画利用这种关系



贝叶斯网络学习

· GaussianNB

GaussianNB假设特征的先验概率为正态分布，即如下式：

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

GaussianNB会根据训练集求出μy和σ2y。 μy为在样本类别中，所有Xi的平均值。σ2y为在样本类别Ck中，所有Xi的方差。

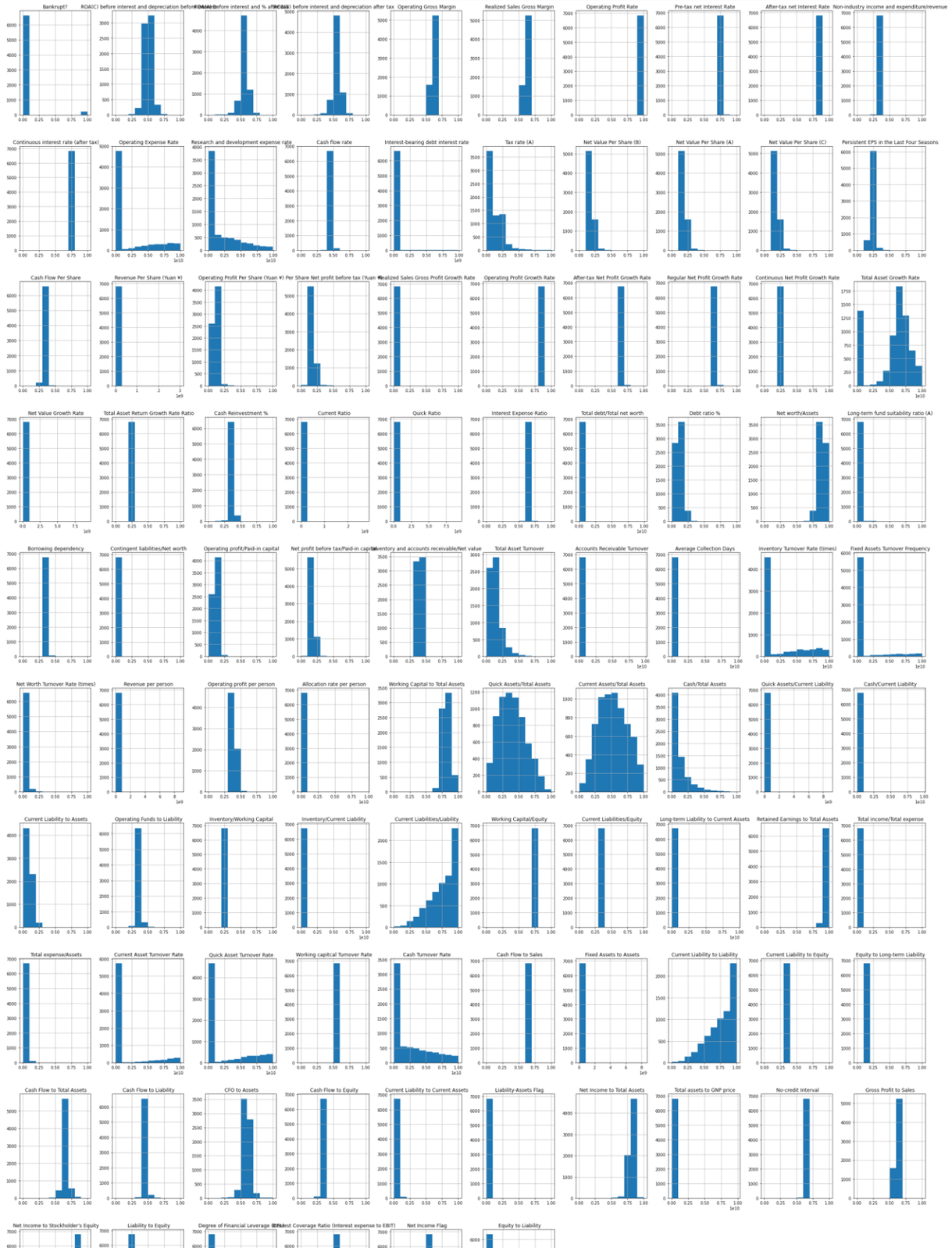
· BernoulliNB

BernoulliNB假设特征的先验概率为二元伯努利分布，即如下式：

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

此时xi只能取值0或者1

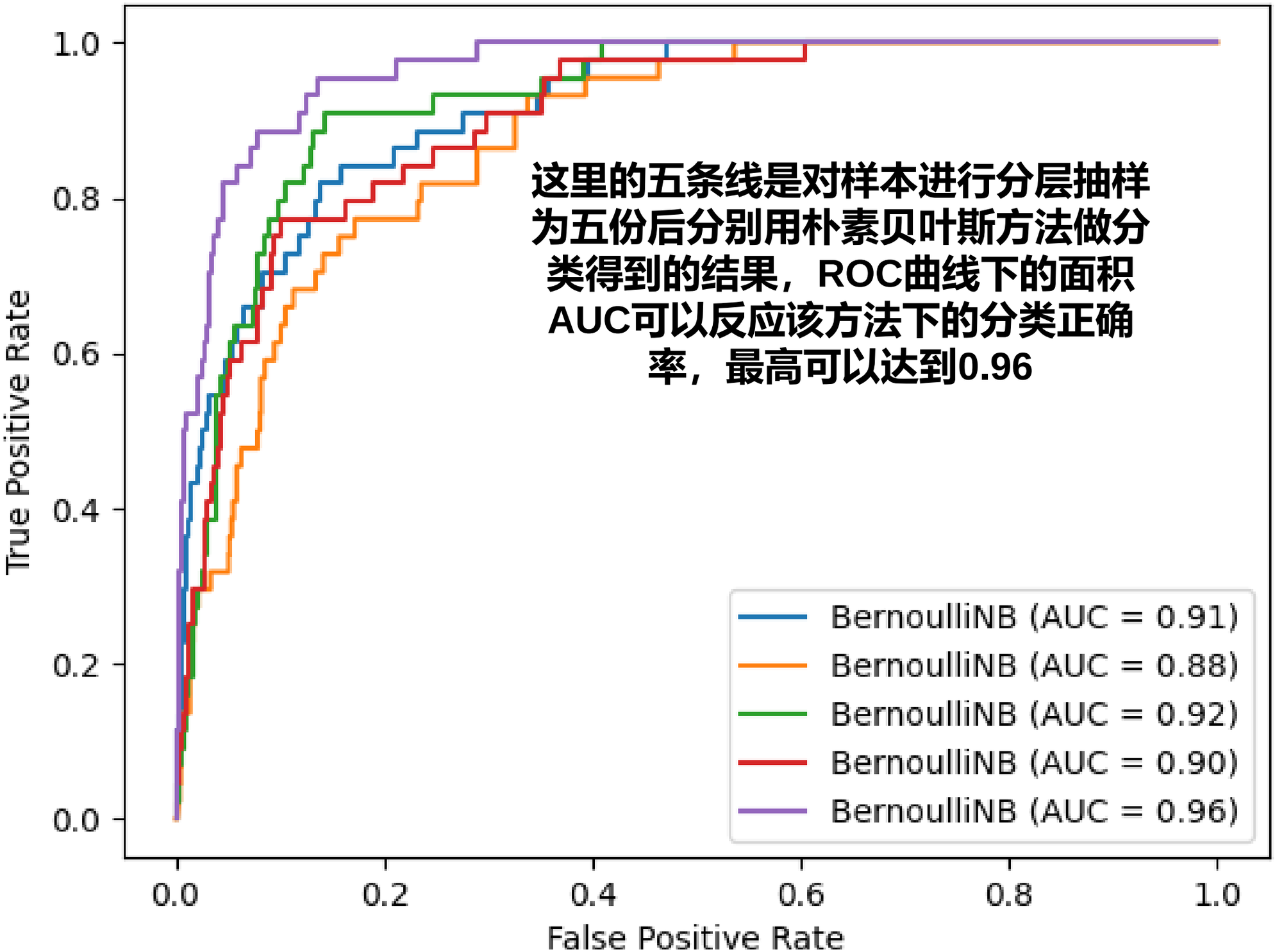
贝叶斯网络学习



这个直方图反映了数据集中各个参数的值的分布，纵轴为频数，我们发现大部分参数拥有相近的取值，所以在GaussianNB外，也尝试了用BernoulliNB对数据集进行处理。

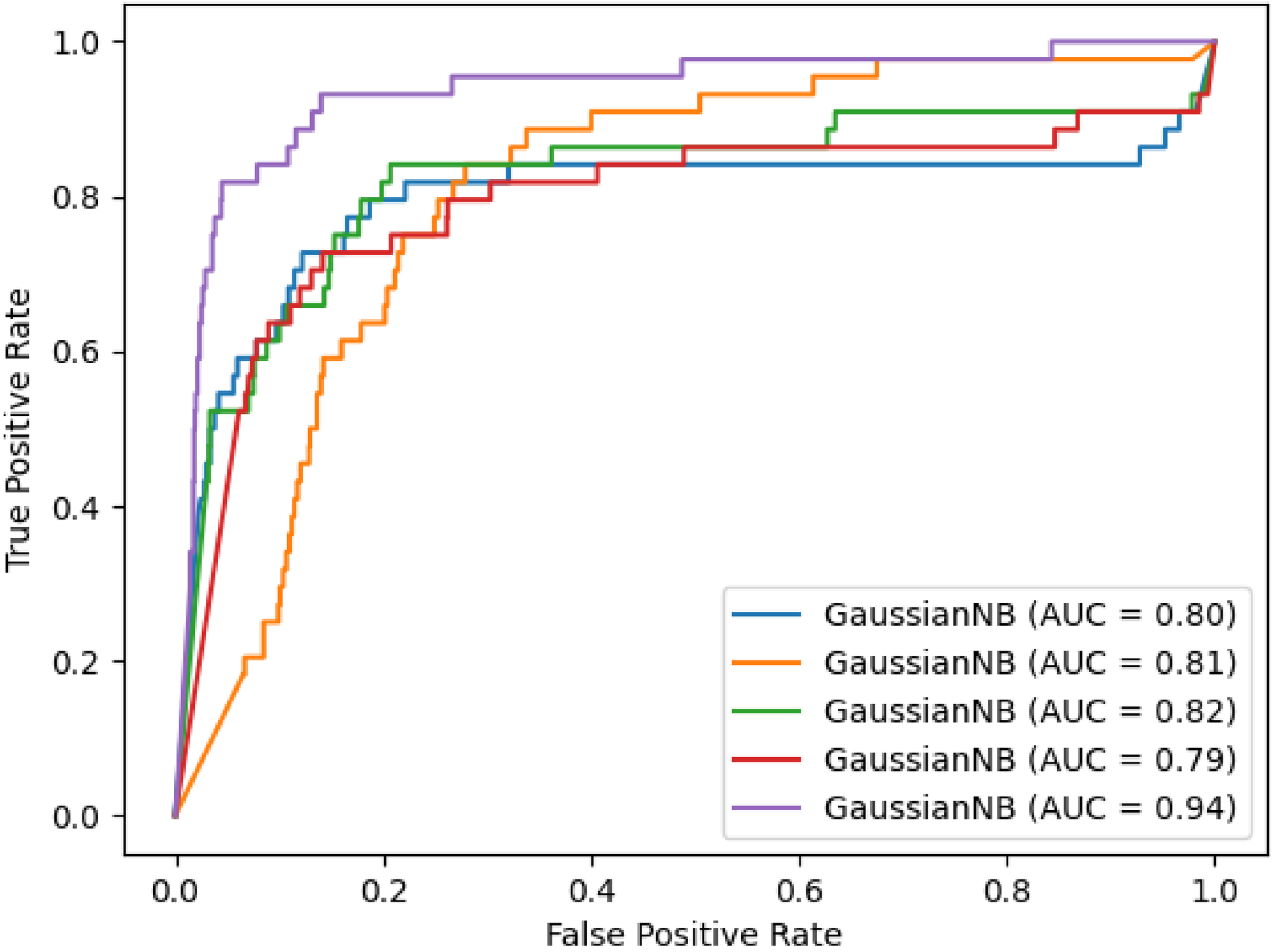
贝叶斯网络学习

- 伯努利分布模型下的朴素贝叶斯分类



贝叶斯网络学习

- 高斯分布模型下的朴素贝叶斯分类



结论：结果对比

- 回顾：accuracy, precision, recall, F1
  - accuracy: 模型预测正确的结果所占的比例

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  - precision: 在被识别为正类别的样本中，确实为正类别的比例是多少

$$\text{Precision} = \frac{TP}{TP + FP}$$

结论：结果对比

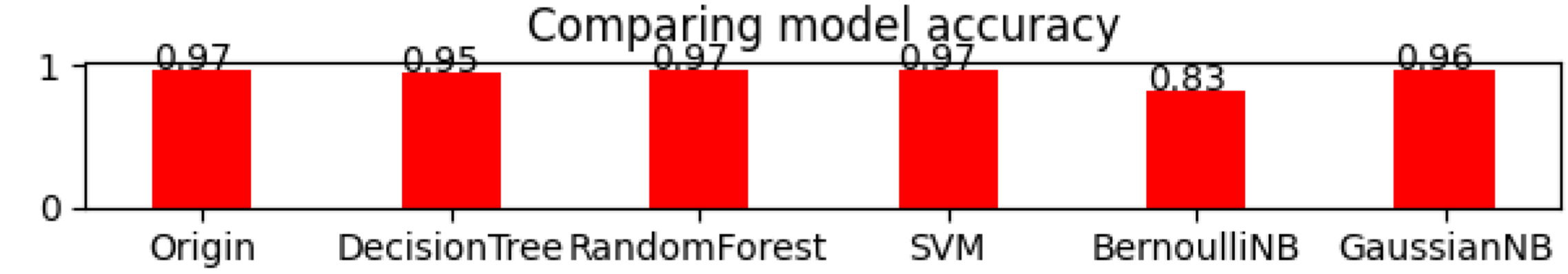
- 回顾：accuracy, precision, recall, F1
  - recall：在所有正类别样本中，被正确识别为正类别的比例是多少

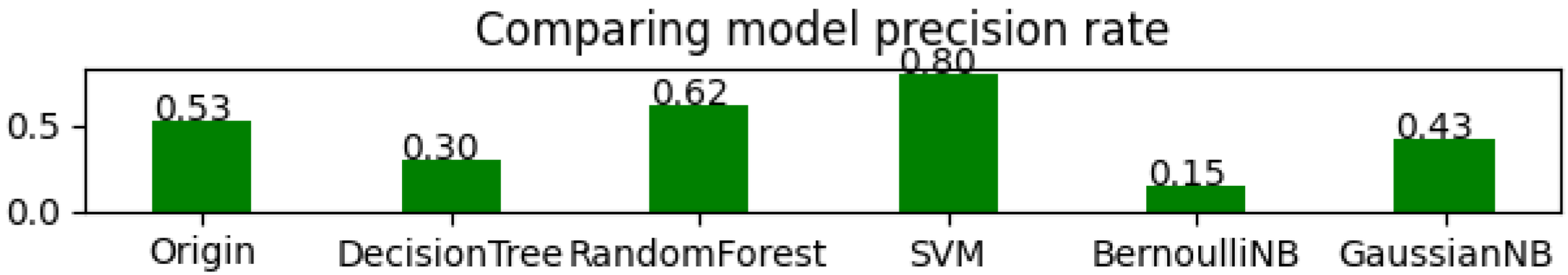
$$\text{Recall} = \frac{TP}{TP + FN}$$

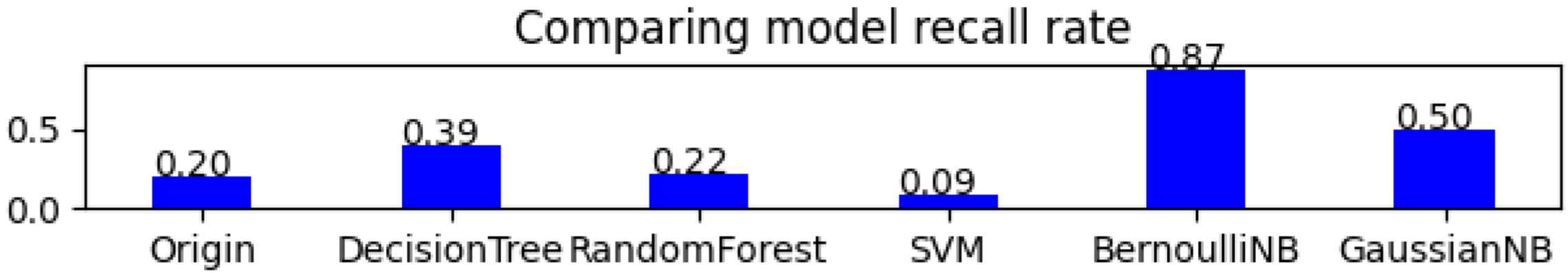
- F1: precision 和 recall 的综合

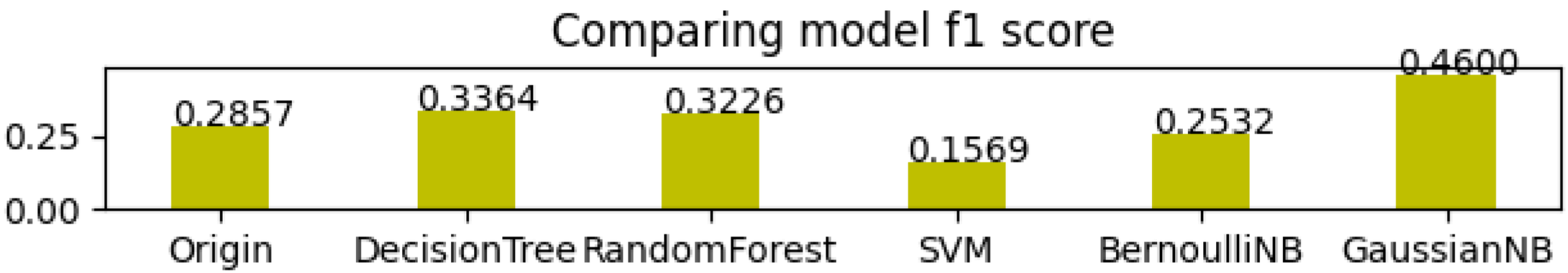
$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}}$$

结论：结果对比


Comparing model accuracy
1
0
0.97
0.95
0.97
0.97
0.83
0.96
Origin
DecisionTree
RandomForest
SVM
BernoulliNB
GaussianNB


Comparing model precision rate
0.5
0.0
0.53
0.30
0.62
0.80
0.15
0.43
Origin
DecisionTree
RandomForest
SVM
BernoulliNB
GaussianNB

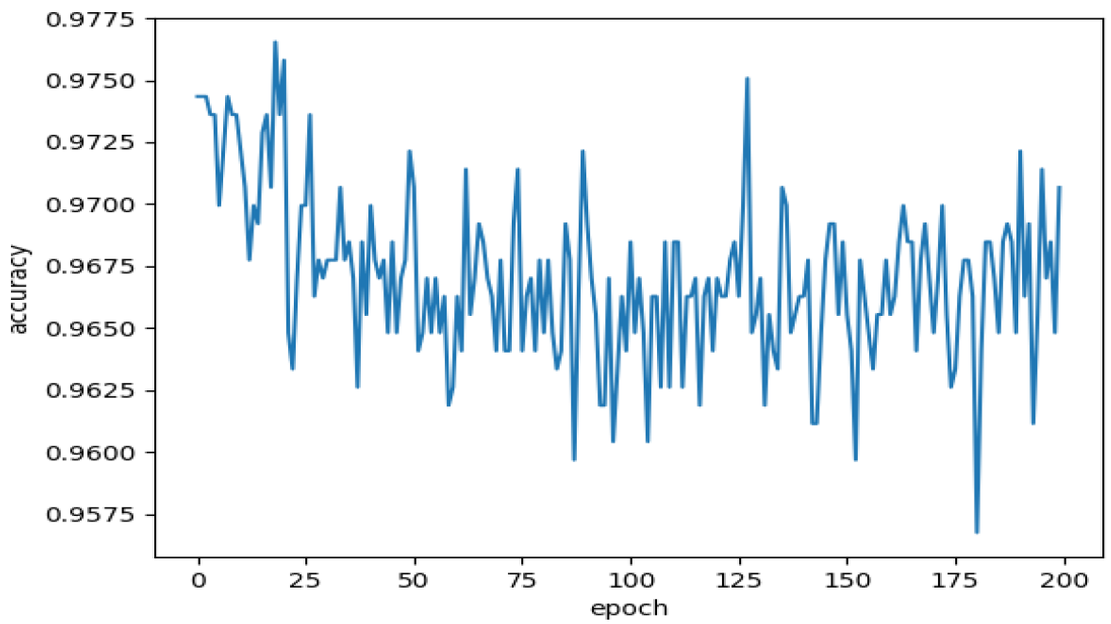

Comparing model recall rate
0.5
0.0
0.20
0.39
0.22
0.09
0.87
0.50
Origin
DecisionTree
RandomForest
SVM
BernoulliNB
GaussianNB

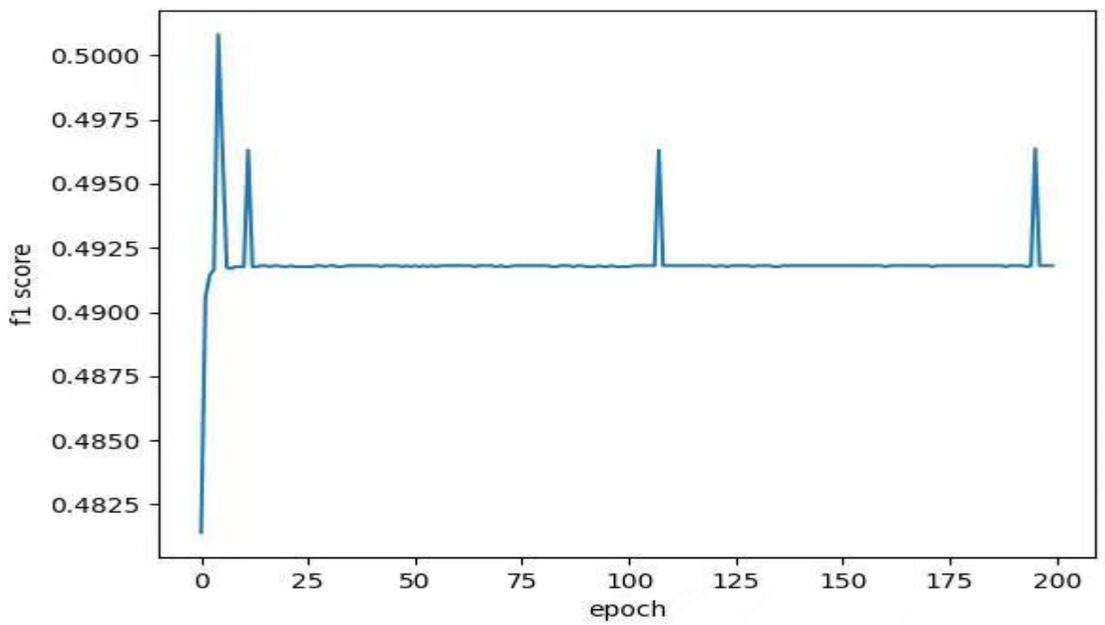

Comparing model f1 score
0.25
0.00
0.2857
0.3364
0.3226
0.1569
0.2532
0.4600
Origin
DecisionTree
RandomForest
SVM
BernoulliNB
GaussianNB

结论：结果对比

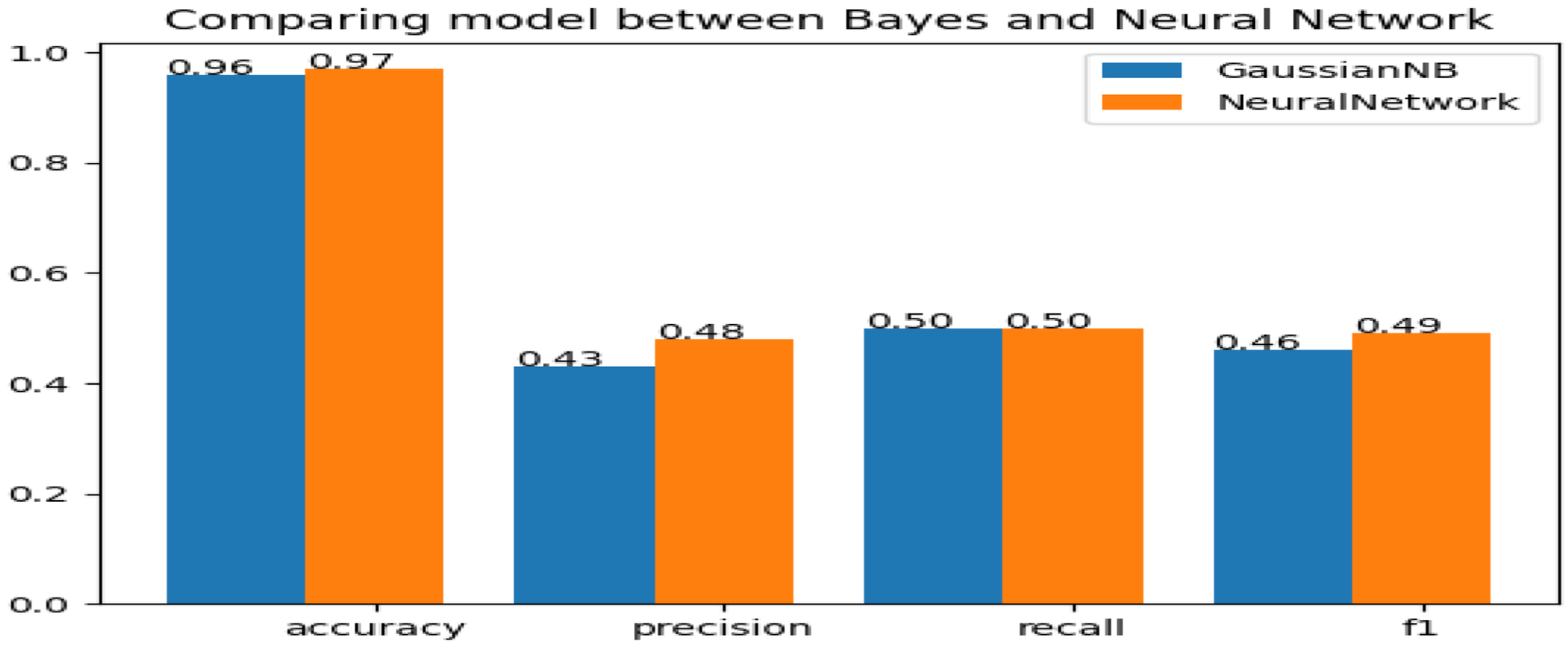
- Origin: 直接在原数据集上训练，图中以随机森林为例
- 其他：数据集经过特征工程和平衡学习后不同模型的结果
- 效果：
  - 可以看出，以随机森林为例，在不改变accuracy的前提下，precision rate, recall rate 和 f1 score 均有不同程度的提升
  - 贝叶斯方法牺牲了一定的accuracy，但在其他三个指标上显著优于其他方法

结论：与简单神经网络的对比





使用三层全连接网络，结果与贝叶斯方法对比：



结论

- 对于简单问题，传统方法不失为一种好的尝试：
  - 可解释性好
  - 易于训练，计算资源占用少，在实际运用中可以快速得到结果

分工

- 数据降维：张柏舟
- 平衡学习：宋铭宇
- 神经网络：鲁琦琨
- 贝叶斯网：柯佳奇
- 模型整合、可视化：周裕涵

参考文献

Balcaen S., Ooghe H. *35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems.* The British Accounting Review, 38 (2006), pp. 63-93

Lin W.-Y., Hu Y.-H., Tsai C.-F. *Machine learning in financial crisis prediction: A survey* IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews, 42 (4) (2012), pp. 421-436

Ohlson J.A. *Financial ratios and the probabilistic prediction of bankruptcy* Journal of Accounting Research, 18 (1980), pp. 109-131

Bredart X. *Financial distress and corporate governance: The impact of board configuration* International Business Research, 7 (3) (2014), pp. 72-80

Guyon I., Elisseeff A. *An introduction to variable and feature selection* Journal of Machine Learning Research, 3 (2003), pp. 1157-1182

Liang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A. (2016) *Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study*. European Journal of Operational Research, vol. 252, no. 2, pp. 561-572

https://blog.csdn.net/u010654299/article/details/103980964

https://zhuanlan.zhihu.com/p/137826761

https://blog.csdn.net/weixin_43329700/article/details/107325026

https://blog.csdn.net/kizgel/article/details/78553009

http://cs229.stanford.edu/notes2020spring/cs229-notes10.pdf