中风患者的发病模式是什么？

成员分工

姓名	学号	分工
刘晓晨	3220200920	选题调研，系统设计，文档编写
刘晓雨	3220200921	相关实验，可视化分析，文档编写
高佳蕊	3520200005	数据预处理，仓库管理，文档编写
吴林涛	3520190042	设计算法，训练模型，模型优化

1. 介绍

1.1 问题背景及意义

心血管疾病(cardiovascular disease, CVD)是世界范围内威胁人类健康的主要疾病，在世界许多国家已占全死因的第一、二位，同时也是 21 世纪中国所面临的主要公共卫生问题之一。据世界卫生组织称：中风，亦称卒中，是全球第二大死亡原因，约占总死亡人数的 11%。中风是一种严重可怕的心脑血管疾病，病情轻微，会导致不同程度的丧失劳动能力，病情严重的会有致残风险，甚至会有生命危险。

由此可见，脑卒中具有高发病率、高致残率、高死亡率的特点，严重危害着大众健康。陈竺等在进行全国第三次死因回顾性调查后发现导致中国居民死亡的首位疾病为脑血管病[1]。2015 年的《中国居民营养与慢性病状况报告》也显示，最近几年中国居民每年因脑血管病死亡的人数多达 200 万，占总死亡人数的 2. 8%[2]。我国 40-74 岁居民首次脑卒中标化发病率平均每年增长 8. 3%。40 岁及以上居民脑卒中标化患病率由 2012 年的 1. 89%上升至 2016 年的 2. 19%[3]。甘勇等总结国内外流行病学调查研究结果后发现，近 20 年全球脑卒中的发病率、死亡率均呈下降趋势，患病率较前稍有增加，而我国脑卒中的发病率、患病率、死亡率均呈上升趋势且较全球平均水平高，目前亟待研究的脑卒中发生的环境因素还不完全清楚，因此我国脑卒中的防控形势依然较为严峻[4]。

事实上，80%的中风是可以预防的。心血管病和脑卒中的发病时间模式可以提供观察病人什么时间进入脑卒中发作的触发时间(启动一系列连锁反应的第一步），以及可能的发生、发展和复发的预防措施。但是，以前的研究已经证明，中国人群与白种人在心血管疾病的发病模式方面存在很大不同。与西方人群相比，中国人群脑血管疾病的发病率和死亡率明显高于冠心病。将西方的心血管疾病的

发病风险预测模型应用到中国人群仍存在争议。

在本工作中，我们试图将一些导致中风的关键指标形象化。这里的数据是从各种年龄组、性别、习惯和与健康有关的问题中抽取的。我们的目的是可视化各种健康和不健康习惯与中风之间的关系，并通过最佳模型和超调参数预测来中风概率。我们的目标是：利用最佳模型和超调参数预测脑卒中的发生概率，建立不同健康和不健康生活习惯与脑卒中的关系。

1.2 数据来源

我们选用来自 kaggle 的数据集[Stroke Prediction Dataset]。该数据集用于根据输入参数(如性别、年龄、各种疾病和吸烟状况)预测患者是否可能患中风。数据中的每一行都提供有关患者的相关信息。

2. 数据预处理

中风预测数据集数据的每一行提供了患者的相关信息，用于根据输入参数预测患者是否有可能患中风。本章节将进行数据属性分类、数据统计、属性分析和数据缺失处理。

首先对数据进行载入并初步查看：

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
7	10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
8	27419	Female	59.0	0	0	Yes	Private	Rural	76.15	NaN	Unknown	1
9	60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1

2.1 参数属性信息

我们首先对中风数据的基本信息进行展示：

```
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   int64
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64
 4   heart_disease      5110 non-null   int64
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
```

可以看出，本数据共 5110 行，即 5110 个样本，但是部分样本 BMI 有缺失。参数属性信息共有如下 12 个类别：

1. ID：参与者的身份号码；
2. 性别：参与者的性别，男、女或其他；
3. 年龄：参与者的年龄，0 至 100 的整数；
4. 高血压：参与者的健康相关参数，有或无；
5. 心脏病：参与者的健康相关参数，有或无；
6. 婚姻状况：参与者的婚姻状况，是或否；
7. 工作类型：参与者工作场所的性质；
8. 居住类型：参与者的居住类型；
9. 平均葡萄糖水平：参与者的健康相关参数，平均葡萄糖水平；
10. BMI 指数：参与者的身体质量指数；
11. 吸烟情况：参与者的习惯性信息，是或否；
12. 中风：参与者是否患有中风，是或否。

我们可以将属性分类为三种：

- 性别、是否已婚、工作类型、居住类型、是否吸烟属于标称特征；
- 高血压、心脏病、中风属于二进制数值特征；
- id、年龄、平均葡萄糖水平、BMI 指数属于连续数值特征。

2.2 数值属性

对于数值属性，我们运用五数概括对其进行分析。

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.00	5110.00	5110.0	5110.00	5110.00	4909.00	5110.00
mean	36517.83	43.23	0.1	0.05	106.15	28.89	0.05
std	21161.72	22.61	0.3	0.23	45.28	7.85	0.22
min	67.00	0.08	0.0	0.00	55.12	10.30	0.00
25%	17741.25	25.00	0.0	0.00	77.24	23.50	0.00
50%	36932.00	45.00	0.0	0.00	91.88	28.10	0.00
75%	54682.00	61.00	0.0	0.00	114.09	33.10	0.00
max	72940.00	82.00	1.0	1.00	271.74	97.60	1.00

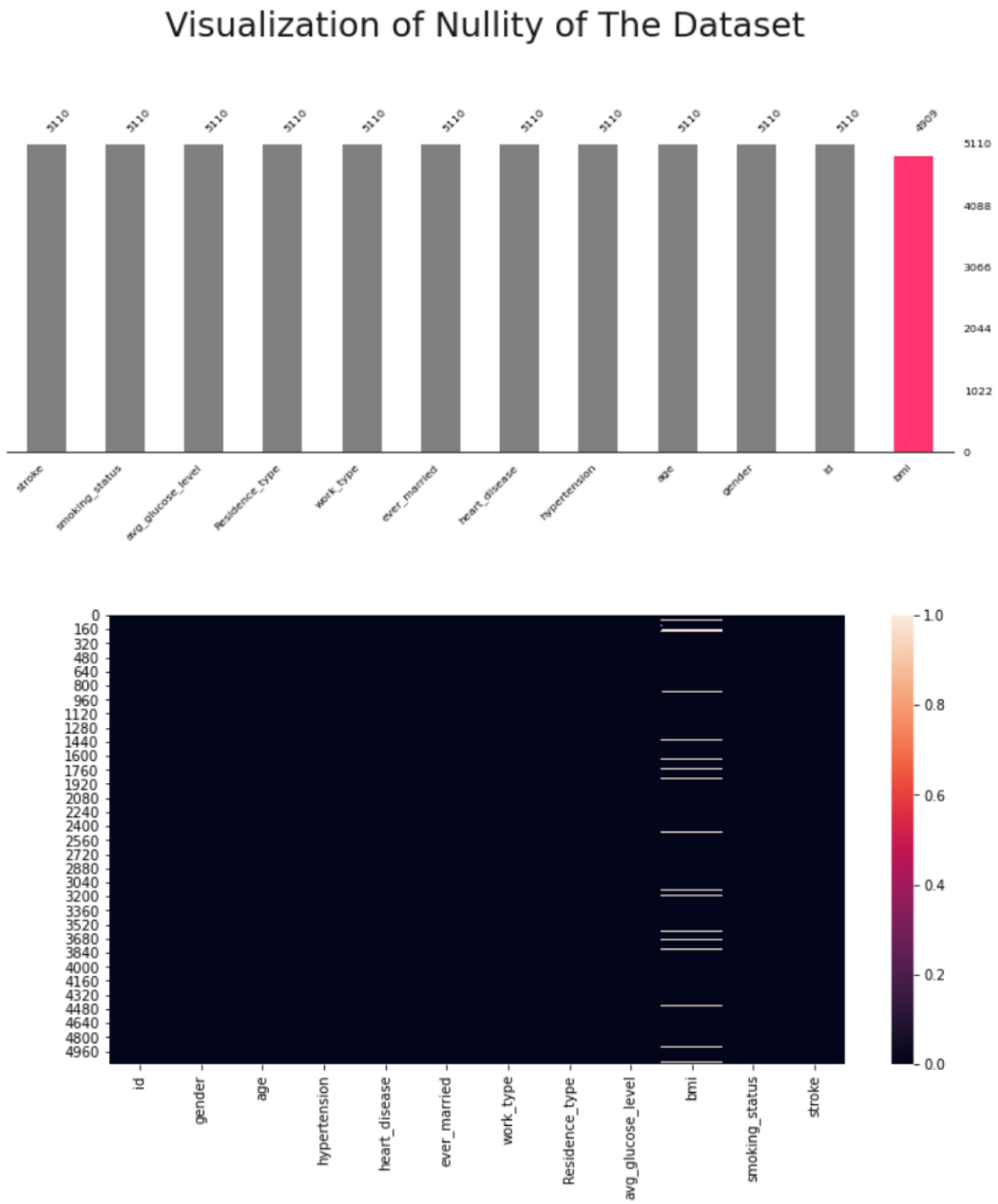
2.3 标称属性

对于标称属性，我们对其进行频数的统计。

	gender	ever_married	work_type	Residence_type	smoking_status
count	5110	5110	5110	5110	5110
unique	3	2	5	2	4
top	Female	Yes	Private	Urban	never smoked
freq	2994	3353	2925	2596	1892

2.4 缺失数据处理

通过下图，我们可以发现 5110 个样本中，只有 bmi 属性有一些缺失值。



缺失人数为 201 人，占比 0.33%。接下来，我们将中位数填入空缺位置。

```
df['bmi'].fillna(df['bmi'].mean(), inplace=True)
```

```
# 补充之后的数据
df.isnull().sum()
```

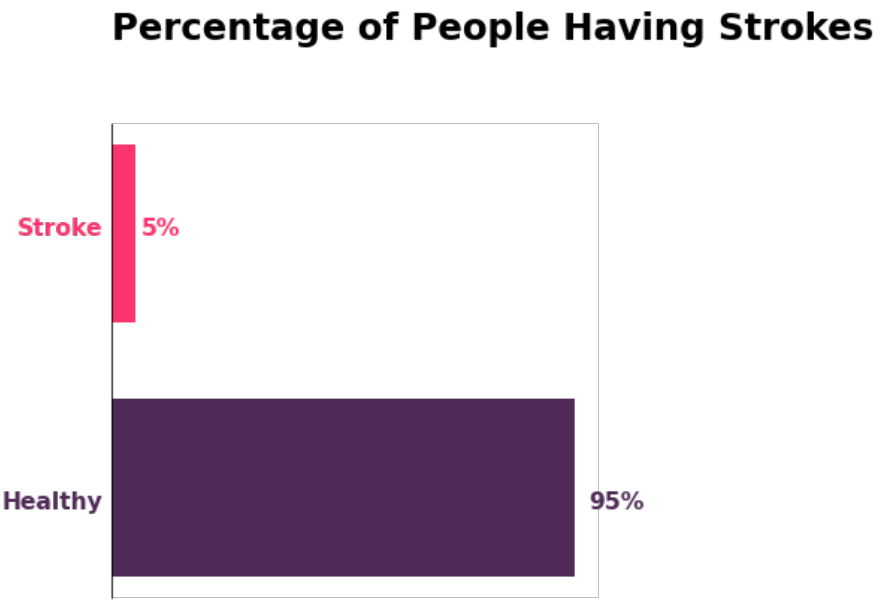
```
id                 0
gender             0
age                0
hypertension       0
heart_disease      0
ever_married       0
work_type          0
Residence_type     0
avg_glucose_level  0
bmi                0
smoking_status     0
stroke             0
dtype: int64
```

这一步完成之后，我们可以获得了完整的数据，之后对 BMI、AGE 和 GLUCOSE 进行范围划分，它们便有数值属性变成了标称属性，更有助于后续的分析。

```
df['bmi_cat'] = pd.cut(df['bmi'], bins = [0, 19, 25,30,10000], labels = ['Underweight', 'Ideal', 'Overweight', 'Obesity'])
df['age_cat'] = pd.cut(df['age'], bins = [0,13,18, 45,60,200], labels = ['Children', 'Teens', 'Adults','Mid Adults','Elderly'])
df['glucose_cat'] = pd.cut(df['avg_glucose_level'], bins = [0,90,160,230,500], labels = ['Low', 'Normal', 'High', 'Very High'])
```

3. 数据初探索

3.1 总体患病比例

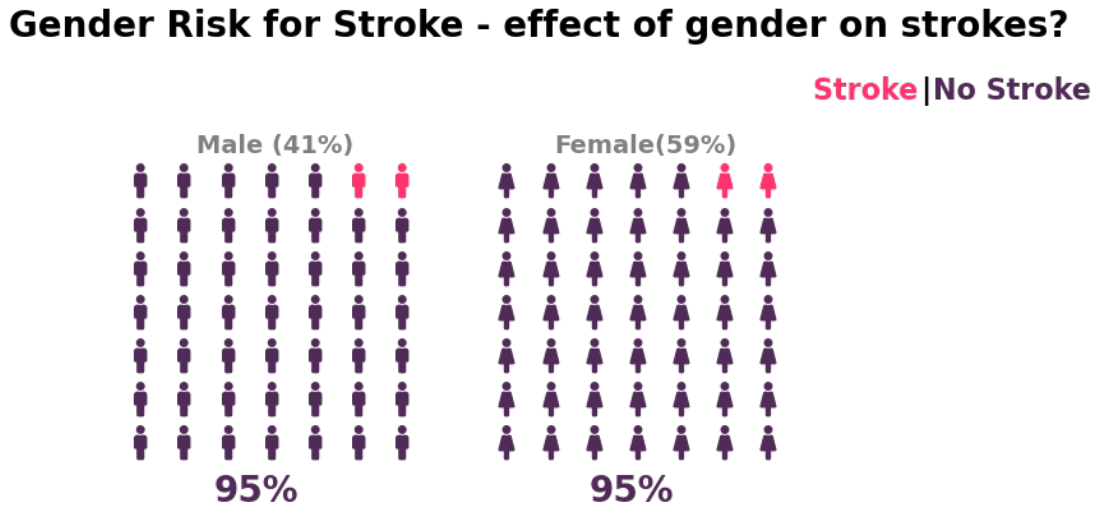


从分布中可以明显看出，根据我们的抽样数据，每 100 个人中就有 5 个人患有中风。 而且，这是一个高度不平衡的数据分布，该分布本身的无效准确度得分为 95%，因此，如果采用任何模型，随机预测准确度可以达到 95%。 所以，在对数据进行建模和训练时，必须进行过采样或欠采样以获得最佳结果。

3.2 分类变量的单变量分析

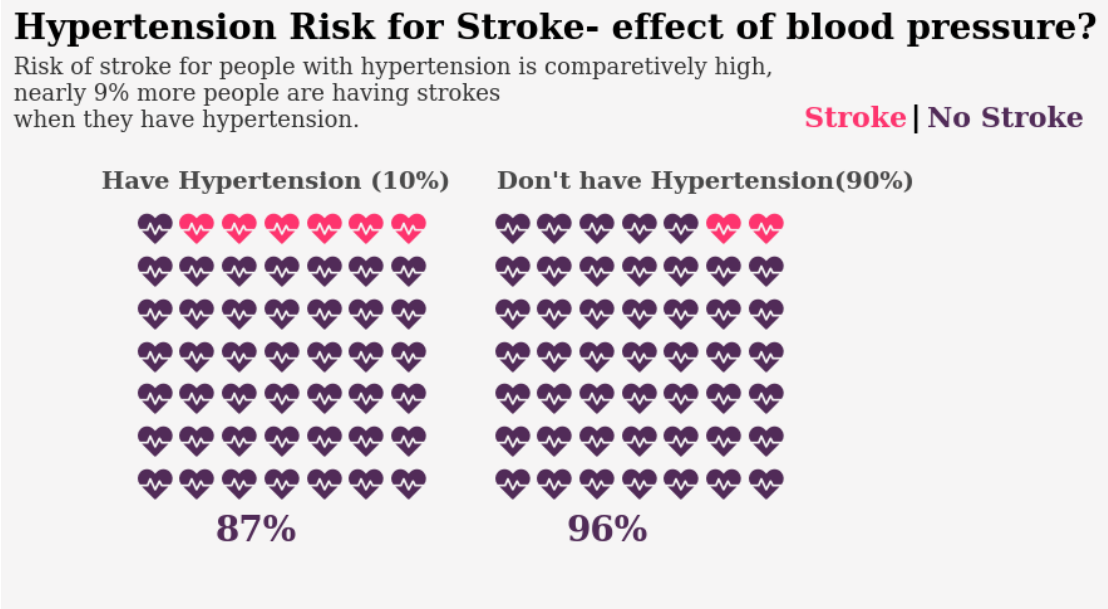
在本节中，我们通过对性别，是否患有高血压，是否患有心脏病这几类单变量做单独分析，得出它们对中风的影响趋势。

3.2.1 对性别的分析



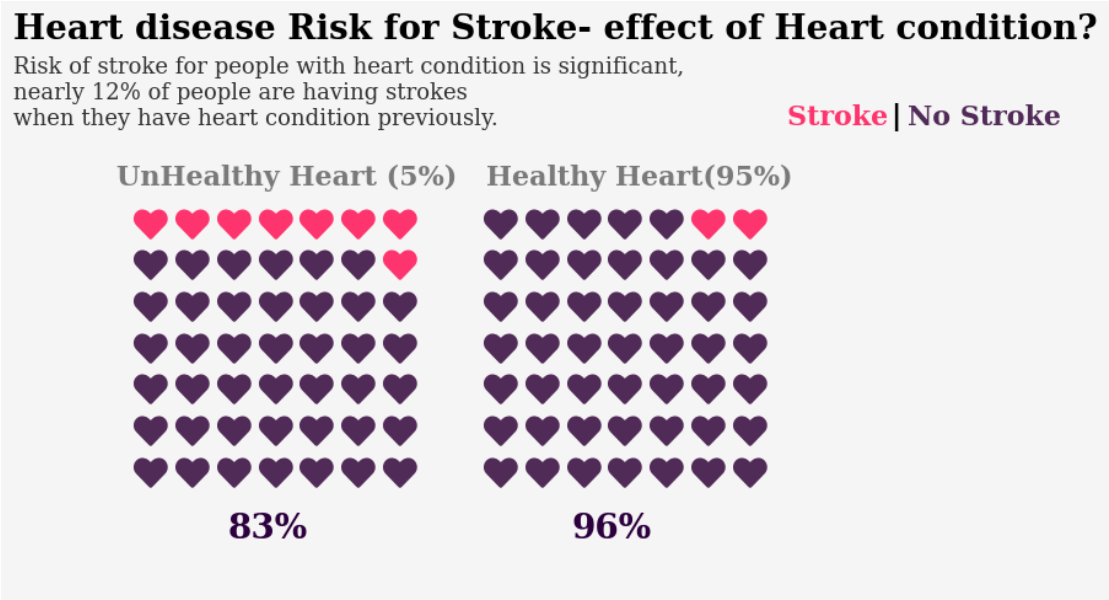
在样本中，男性和女性的人数是大约相等的，我们将样本分为男性样本和女性样本，并统计两份样本中患有中风的人数的。结果如上图所示，从中我们可以看出，男性和女性中风的风险几乎是相同的，都在 5%左右。

3.2.2 对患有高血压的分析



从中我们可以看出，患有高血压的人群患中风的比例为 13%，而不患病的人群只有 4%。因此我们可以得出结论：患有高血压的人群比不患高血压的人群更容易中风，这提醒我们两种疾病之间可能会存在某种联系。

3.2.3 对患有心脏病的分析



与上小节结果类似，从图中我们可以看出，患有心脏病的人群比不患病的人群更容易中风。

4. 可视化数据平衡与数据采样技术

4.1 SMOTE 简介

SMOTE（Synthetic Minority Oversampling Technique），合成少数类过采样技术. 它是基于随机过采样算法的一种改进方案，由于随机过采样采取简单复制样本的策略来增加少数类样本，这样容易产生模型过拟合的问题，即使得模型学习到的信息过于特别(Specific)而不够泛化(General)，SMOTE 算法的基本思想是对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中，算法流程如下。

(1)对于少数类中每一个样本 x，以欧氏距离为标准计算它到少数类样本集中所有样本的距离，得到其 k 近邻。

(2)根据样本不平衡比例设置一个采样比例以确定采样倍率 N，对于每一个少数类样本 x，从其 k 近邻中随机选择若干个样本，假设选择的近邻为 xn。

(3)对于每一个随机选出的近邻 xn，分别与原样本按照如下的公式构建新的样本。

4.2 实现

我们选择 SMOTE 过采样数据进行建模，因为该技术生成的数据点的个数比例相等。

```
Inverse of Null Accuracy:  0.04872798434442227
Null Accuracy:  0.9512720156555773
```

计算得到零精度达到了 95%，零精度的逆为 4.9%。

5. 预测模型建立及训练

5.1 数据集划分

首先将数据集划分为训练集和测试集：

```
Shape of Training features: (7284, 24)
Shape of Training targets: (7284,)
Shape of Testing features: (1278, 24)
Shape of Testing targets: (1278,)
```

然后计算得到当前数据的零精度为 95%

```
Null Accuracy Score for Current Data is 0.95
```

5.2 评价指标

（1）F1 分数（F1 Score），是统计学中用来衡量二分类模型精确度的一种指标。它同时兼顾了分类模型的精确率和召回率。F1 分数可以看作是模型精确率和召回率的一种调和平均，它的最大值是 1，最小值是 0。

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

（2）AUC（Area Under Curve）被定义为 ROC 曲线下与坐标轴围成的面积，显然这个面积的数值不会大于 1。又由于 ROC 曲线一般都处于 y=x 这条直线的上方，所以 AUC 的取值范围在 0.5 和 1 之间。AUC 越接近 1.0，检测方法真实性越高;等于 0.5 时，则真实性最低，无应用价值。

其中,ROC 曲线的横坐标是伪阳性率(也叫假正类率,False Positive Rate)，纵坐标是真阳性率（真正类率，True Positive Rate)，相应的还有真阴性率（真负类率，True Negative Rate）和伪阴性率（假负类率，False Negative Rate)。这四类指标的计算方法如下：

- 伪阳性率（FPR)：判定为正例却不是真正例的概率，即真负例中判为正例的概率
- 真阳性率（TPR)：判定为正例也是真正例的概率，即真正例中判为正例

的概率（也即正例召回率）

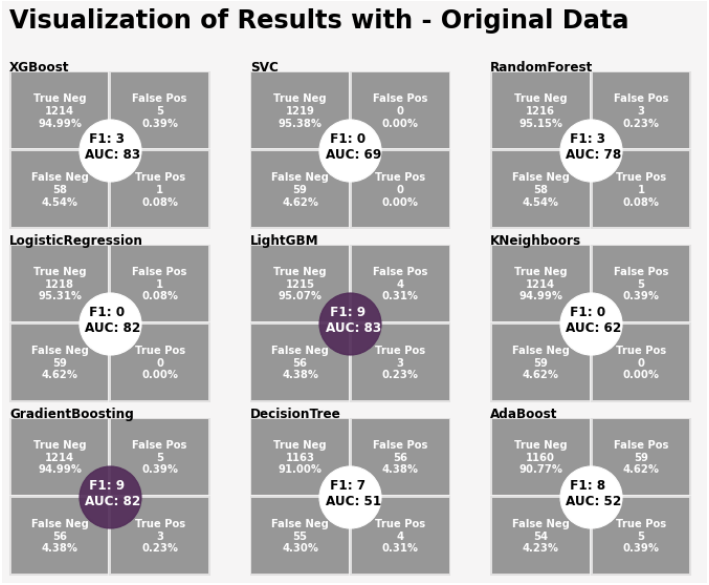
- 伪阴性率（FNR）：判定为负例却不是真负例的概率，即真正例中判为负例的概率。
- 真阴性率（TNR）：判定为负例也是真负例的概率，即真负例中判为负例的概率。

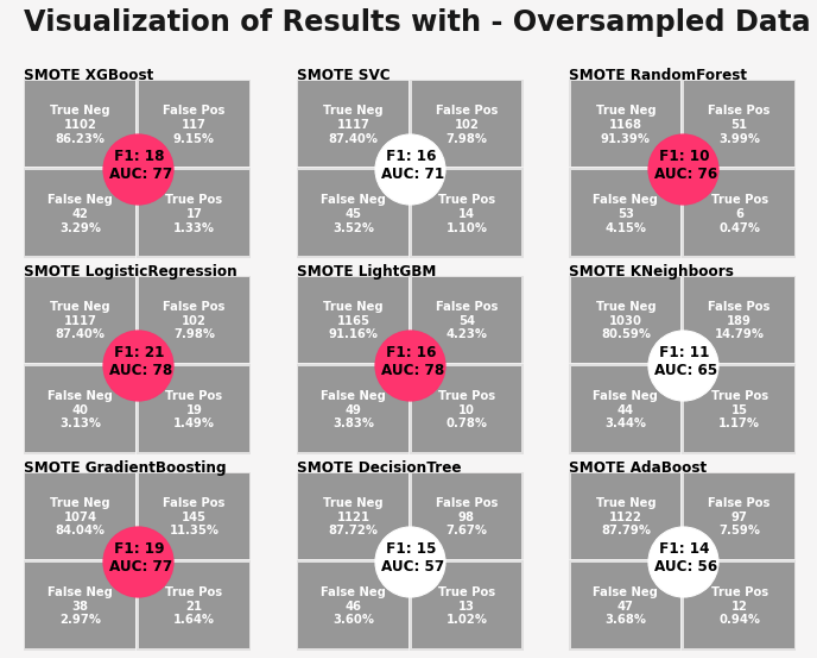
5.3 构建模型

最后构建预测模型并训练，本节采用 SVC、DecisionTree、AdaBoost、RandomForest、GradientBoosting、KNeighboors、LogisticRegression、XGBoost、LightGBM 这 9 种分类模型分别对原始数据和过采样数据进行处理，得到它们的 F1 得分和 AUC。

5.4 可视化分析

对它们的实验结果进行可视化处理以及对比分析，如下图所示。





上图分为两个部分，第一部分是原始数据的可视化结果，第二部分是过采样数据的可视化结果,从可视化中可以清楚地看到，过采样数据比原始数据具有更好的预测分数。原始数据中只有采用 LightGBM 和 GradientBoosting 这两种分类方法的结果 F1 分数大于 5 且 AUC 达到了 75%以上，而过采样数据中 9 种分类方法的 F1 分数都在 10 以上且有 XGBoost、RandomForest、LogisticRegression、LightGBM 和 GradientBoosting5 种分类方法处理过的结果对比原始数据，它们的 F1 分数要高很多并且 AUC 达到了 75%以上。

6. 总结

经过我们的实验发现，对于原始数据来说 LightGBM 和 GradientBoosting 这两种预测模型预测的准确率较高，性能较好；对于过采样数据来说 LogisticRegression 方法取得的效果最好。并且我们的实验还表明使用过采样数据比使用原始数据对于所有预测模型来说性能都有所提升，效果更好。