回到根源！

这是我在Kaggle旅程中发布的第二个内核，我想对它进行一些更改，以使其对Kaggle社区更具吸引力。如果你喜欢这个内核，请upvote。每当我从学校获得空闲时间，我都会对这个内核进行更新，希望你喜欢它。让我们开始！

银行营销数据集 - 智能定位：

营销介绍：

公司为客户创造价值并建立牢固的客户关系以便从客户那里获取价值的过程。

科特勒和阿姆斯特朗（2010年）。

营销活动的特点是关注客户需求和整体满意度。然而，有不同的变量决定营销活动是否会成功。在进行营销活动时，我们需要考虑某些变量。

4 P：

1）人口部分：营销活动将针对哪一部分人群进行解决？为什么？营销活动的这一方面非常重要，因为它将告诉哪些人群最有可能收到营销活动的信息。

2）到达客户所在地的分销渠道：实施最有效的策略，以充分利用此营销活动。我们应该解决哪一部分人口问题？我们应该使用哪种工具来传达信息？ （例如：电话，广播，电视，社交媒体等）

3）价格：向潜在客户提供的最优价格是多少？ （在银行营销活动的情况下，这是没有必要的，因为银行的主要利益是潜在客户开设转账账户，以使银行的运营活动继续运行。）

4）促销策略：这是战略实施的方式以及潜在客户如何解决。这应该是营销活动分析的最后一部分，因为必须对以前的活动进行深入分析（如果可能），以便从以前的错误中吸取教训并确定如何使营销活动更加有效。

关于这个内核：

我知道这是一个众所周知的数据集，因为它来自UCI机器学习库。但是，我相信您可以看到一些有趣的见解，您可以将它们集成到您自己的数据分析中。总而言之，Kaggle意在向别人学习，我希望这个例子很适合你。

请随意将此内核用于您的项目，这将是我的荣幸！

此外，我愿意接受新的想法和我可以改进的东西，以使这个内核更好！开放构建批评！最后，我要特别感谢Randy Lao和他着名的Predicting Employee Kernelover。他的内核给了我不同的想法，我应该如何处理数据集的分析。

另外，我想赞扬这个stackoverflow帖子，它帮助我改变了Facetgrids的传说名称。

https://stackoverflow.com/questions/45201514/edit-seaborn-plot-figure-legend

如果您正在努力解决同样的问题，请查看它。

新更新：

确定最有可能开立定期存款账户的样本群体中的群集。

什么是定期存款？

定期存款是银行或金融机构以固定利率（通常优于仅开户存款账户）提供的存款，其中您的资金将在特定到期时间返还。有关定期存款的更多信息，请单击Investopedia上的链接：https：//www.investopedia.com/terms/t/termdeposit.asp

大纲：

A.属性描述

I.银行客户数据

II。与当前广告系列的最后一次联系相关

III。其他属性

B.构建数据：

一，数据的整体分析

II。数据结构和转换

C.探索性数据分析（EDA）

I.接受与被拒绝的定期存款

II。分布图

D.分析的不同方面：

I.营销活动的月份

II。 Seasonalities

III。潜在客户的呼叫数量

IV。潜在客户的年龄

V.导致更多定期存款的职业类型

E.影响潜在客户决策的相关性。一，我们的相关矩阵分析

II。平衡类别与住房贷款

III。 H.Loans与定期存款之间的负相关关系

F.分类模型

一，简介

II。分层抽样

III。分类模型

IV。混乱矩阵

V.精度和回忆曲线

VI。特征重要性决策树C.

G.下一个竞选战略

一，银行应考虑的行动

A.属性描述：

输入变量：

艾。银行客户数据：

1 - 年龄:(数字）

2 - 工作：工作类型（分类：'管理员'，'蓝领'，'企业家'，'女佣'，'管理'，'退休'，'自雇'，'服务'，'学生' '技术人员'， '失业'， '未知'）

3 - 婚姻：婚姻状况（分类：'离婚'，'已婚'，'单身'，'未知';注：'离婚'是指离婚或丧偶）

4 - 教育:(分类：小学，中学，大专和未知）

5 - 默认值：默认信用额度？ （分类：'不'，'是'，'未知'）

6 - 住房：有住房贷款吗？ （分类：'不'，'是'，'未知'）

7 - 贷款：有个人贷款吗？ （分类：'不'，'是'，'未知'）

8 - 平衡：个人的平衡。

AII。与当前广告系列的最后一次联系相关：

8 - 联系方式：联系沟通类型（分类：'蜂窝'，'电话'）

9个月：上一个联系月份（分类：'jan'，'feb'，'mar'，...，'nov'，'dec'）

10天：一周的最后一个联系日（分类：'mon'，'tue'，'wed'，'thu'，'fri'）

11 - 持续时间：上次联系持续时间，以秒为单位（数字）。重要提示：此属性会严重影响输出目标（例如，如果持续时间= 0，则y ='否'）。然而，在执行呼叫之前不知道持续时间。此外，在通话结束后，显然已知y。因此，此输入仅应包括在基准目的中，如果打算采用现实的预测模型，则应将其丢弃。

AIII。其他属性：

12 - 广告系列：此广告系列期间和此客户端执行的联系人数量（数字，包括最后一次联系）

13 - pdays：从上一个广告系列上次联系客户端之后经过的天数（数字; 999表示之前未联系过客户）

14 - 上一个：此广告系列之前和此客户端执行的联系人数量（数字）

15 - poutcome：上一次营销活动的结果（分类：'失败'，'不存在'，'成功'）

输出变量（所需目标）：

21 - y - 客户是否订购了定期存款？ （二进制：'是'，'不'）

探索基础知识

摘要：

MeanEge年龄大约41岁。 （最低：18岁，最高：95岁。）

平均余额为1,528。但是，标准偏差（标准偏差）是一个很高的数字，因此我们可以通过此理解，平衡在整个数据集中分布很大。

由于数据信息表明最好放弃持续时间列，因为持续时间与潜在客户是否会购买定期存款高度相关。此外，在向潜在客户端进行呼叫之后获得持续时间，因此如果目标客户端从未接收过呼叫，则该功能不是那么有用。持续时间与开立定期存款高度相关的原因是银行与目标客户谈判的次数越多，目标客户开立定期存款的可能性就越高，因为持续时间越长意味着潜在客户的更高利益（承诺） 。

注意：我们可以从描述性数据集中获得太多的见解，因为我们的大多数描述性数据不在“数字”列中，而是在“分类列”中。

幸运的是，没有遗漏的值。如果缺少值，我们将不得不用中位数，均值或模式填充它们。我倾向于使用中位数，但在这种情况下，不需要填写任何缺失值。这肯定会让我们的工作更轻松！

按职业分析：

职业数量：管理是此数据集中更为普遍的职业。

•按职业分类：正如预期的那样，退休人员的年龄中位数最高，而学生最低。

职业平衡：管理层和退休人员是账户余额最高的人。

婚姻状况

在这个分析中，除了大多数离婚的人都破产之外，我们没有发现任何重要的见解。难怪因为他们必须拆分金融资产！尽管如此，由于没有找到进一步的见解，我们将继续将婚姻状况与教育状况分开。让我们看看我们是否可以在样本群体中找到其他人群。

聚集婚姻状况和教育：

婚姻状况：如前所述，离婚的影响对个人的平衡产生重大影响。

教育：教育水平也对潜在客户的平衡数量产生重大影响。

贷款：潜在客户是否有先前的贷款对他或她的余额有重大影响。

活动时长：

Campaign持续时间：嗯，我们看到持续时间与定期存款具有高度相关性意味着持续时间越长，客户开立定期存款的可能性就越大。

平均活动持续时间：平均活动持续时间为374.76，让我们看看高于此平均值的客户是否更有可能开立定期存款。

持续时间：超过持续时间状态的人更有可能开立定期存款。持续时间高于平均水平的群体的78％开设定期存款，而低于平均32％的群体开立定期存款账户。这告诉我们，针对属于上述平均类别的个人是一个好主意。

分类模型：

分层抽样：

分层抽样：在开发用于回归或分类的模型时，这是一个经常被忽略的重要概念。请记住，为了避免过度拟合我们的数据，我们必须实施交叉验证，但是，我们必须确保至少对我们的标签影响最大的功能（潜在客户是否会开立定期存款）是平均分配。这是什么意思？

个人贷款：

例如，拥有个人贷款是决定潜在客户是否会开立定期存款的重要特征。要确认它对最终输出有很大影响，您可以检查上面的相关矩阵，您可以看到它与开立存款有-11％的相关性。在我们的列车和测试数据中实施分层抽样之前我们应该采取哪些步骤？

1）我们需要了解我们的数据是如何分布的。

2）在通知贷款栏中含有87％的“否”（没有个人贷款）和13％的“是”（有个人贷款）。

3）我们希望确保我们的训练和测试集包含相同比例的87％“no”和13％“yes”。“分层抽样：在开发模型时经常会错过一个重要的概念，无论是回归还是记住，为了避免过度拟合我们的数据，我们必须实施交叉验证，但是，我们必须确保至少对我们的标签影响最大的特征（潜在客户是否会开立定期存款） ）是平均分配的。这是什么意思？

个人贷款：

例如，拥有个人贷款是决定潜在客户是否会开立定期存款的重要特征。要确认它对最终输出有很大影响，您可以检查上面的相关矩阵，您可以看到它与开立存款有-11％的相关性。在我们的列车和测试数据中实施分层抽样之前我们应该采取哪些步骤？

1）我们需要了解我们的数据是如何分布的。

2）在通知贷款栏中含有87％的“否”（没有个人贷款）和13％的“是”（有个人贷款）。

3）我们希望确保我们的培训和测试集包含87％“no”和13％“yes”的相同比率。

避免过度拟合：

过度拟合的简要说明？

这是建模算法中的一个错误，它考虑了拟合过程中的随机噪声而不是模式本身。你可以看到，当模型在训练集中得到一个很棒的分数时会发生这种情况，但是当我们使用测试集（模型的未知数据）时，我们会得到一个糟糕的分数。这很可能是因为数据过度拟合（考虑到我们模式中的随机噪声）。我们希望我们的模型做的是采用数据的整体模式，以便正确地分类潜在客户是否会定期存款。在上面的示例中，决策树分类器和随机森林分类器很可能过度拟合，因为它们都给出了近乎完美的分数（100％和99％）准确度分数。

我们怎样才能避免过度拟合？

避免过度拟合的最佳选择是使用交叉验证。参加训练测试并将其分开。例如，如果我们将其拆分为3，则将2/3的数据或66％用于培训，并且将使用1/3 33％或测试，我们将进行三次测试过程。该算法将遍历所有训练和测试集，其主要目的是获取数据的整体模式。

混淆矩阵：

混淆矩阵的见解：

混淆矩阵的主要目的是了解我们的模型在对潜在客户进行分类时的表现，这些客户可能会被定期存款。我们将在混淆矩阵中看到四个术语：真阳性，假阳性，真阴性和假阴性。

正面/负面：类别类型（标签）[“否”，“是”]真/假：模型正确或错误地分类。

真阴性（左上方）：这是“不”类或潜在客户的正确分类数，不愿意扣押定期存款。

假阴性（右上方）：这是“不”类或不愿意承认定期存款的潜在客户的错误分类数。

误报（左下方）：这是“是”类或愿意承保定期存款的潜在客户的错误分类数。

真阳性（右下方）：这是“是”类的正确分类数或愿意承认定期存款的潜在客户。

精度和召回：

回想一下：数据集的标签列中的“是”总数。那么我们的模型检测到多少个“是”标签。

精度：意味着我们的模型预测实际标签是否为“是”。

召回精度权衡：

随着精度的提高，召回率降低，反之亦然。例如，如果我们将精度从30％提高到60％，模型就会选择模型认为60％肯定的预测。如果存在模型认为58％可能是潜在客户的实例，该客户将扣押定期存款，那么该模型将其归类为“否”。然而，那个实例实际上是“是”（潜在客户确实接受定期存款。）这就是为什么模型越有可能错过实际上是“是”的实例！

哪些特征会影响定期存款的结果？

DecisionTreeClassifier：

我们的分类器最重要的三个功能是\*\*持续时间（销售代表与潜在客户之间的对话需要多长时间），联系人（同一营销活动中潜在客户的联系人数），月份（月份）今年）。

GradientBoosting分类器胜了！

Gradient Boosting分类器是预测潜在客户是否会定期支付定期存款的最佳模型。精度84％！

银行应考虑采取什么行动？

下一次营销活动的解决方案（结论）：

1）营销活动的月份：我们看到营销活动的最高月份是5月份。然而，这是潜在客户倾向于拒绝定期存款报价的月份（最低有效税率：-34.49％）。对于下一次营销活动，银行明智地将营销活动集中在3月，9月，10月和12月。 （12月应该正在考虑中，因为它是营销活动最少的月份，可能有12个月最低的原因。）

2）季节性：潜在客户选择在秋冬季节期间使用定期存款。下一个营销活动应该关注这些季节的活动。

3）活动呼叫：应实施一项政策，规定不应对同一潜在客户应用不超过3次呼叫，以节省获取新潜在客户的时间和精力。请记住，我们称之为同一潜在客户的次数越多，他或她可能会拒绝开立定期存款。

4）年龄类别：银行的下一个营销活动应针对年龄在20岁或以下，60岁或以上的潜在客户。最年轻的类别有60％的机会可以存入定期存款，而最年轻的类别有76％的机会可能会被定期存款。如果下一次活动中银行解决了这两个类别，那么将会增加更多定期存款的可能性。

5）职业：毫不奇怪，学生或退休的潜在客户最有可能被定期存款。退休人员往往有更多的定期存款，以通过利息支付获得一些现金。请记住，定期存款是短期贷款，其中个人（在这种情况下是退休人员）同意在个人与金融机构商定的特定日期之前不从银行提取现金。在那之后，个人获得了资本回报，并对贷款产生了兴趣。退休人员往往不会花大钱现金，所以他们更愿意通过贷款给金融机构来投入现金。学生是另一组曾经被定期存款的人。

6）房屋贷款和余额：低余额和无余额类别的潜在客户比平均和高余额类别的人更有可能获得房屋贷款。拥有房屋贷款意味着什么？这意味着潜在客户有经济上的妥协来偿还其房屋贷款，因此，他或她没有现金可以存入定期存款账户。但是，我们发现平均和高余额的潜在客户不太可能获得房屋贷款，因此更有可能开立定期存款。最后，下一次营销活动应该关注平均和高余额的个人，以增加怀疑定期存款的可能性。

7）在通话过程中制定问卷：由于通话时间是与潜在客户是否会开立定期存款最正相关的特征，通过在通话期间为潜在客户提供有趣的问卷，会话长度可能会增加。当然，这并不能保证我们潜在的客户会被定期存款！尽管如此，我们并没有通过实施一项策略来解决任何问题，这种策略会提高潜在客户的参与度，从而导致增加定期存款的可能性，从而提高银行优惠的下一次营销活动的效率。

8）针对持续时间较长的人（375以上）：针对持续时间高于平均水平的目标群体，该目标群体很可能会开设定期存款账户。该集团开立定期存款账户的可能性为78％，相当高。这将使下一次营销活动的成功率非常高。

通过结合所有这些策略并简化下一个广告系列应该解决的市场受众，银行的下一个营销活动可能比当前的更有效。

参考文献：

1）AurélienGéron使用Scikit-Learn和TensorFlow进行动手机器学习。

2）特别感谢Ahmet Taspinar。 （关于一系列功能的见解）链接：http：//ataspinar.com/2017/05/26/classification-with-scikit-learn/

3）特别感谢Randy Lao和他的Predicting Employee KernelOver工作。链接：https://www.kaggle.com/randylaosat/predicting-employee-kernelover