问1：你会在时间序列数据集上使用什么交叉验证技术？是用k倍或LOOCV？

　　答：都不是。对于时间序列问题，k倍可能会很麻烦，因为第4年或第5年的一些模式有可能跟第3年的不同，而对数据集的重复采样会将分离这些趋势，而我们最终可能只是需要对过去几年的进行验证，这就不能用这种方法了。相反，我们可以采用如下所示的5倍正向链接策略：

　　fold 1 : training [1], test [2]

　　fold 2 : training [1 2], test [3]

　　fold 3 : training [1 2 3], test [4]

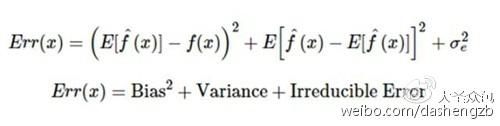
　　fold 4 : training [1 2 3 4], test [5]

　　fold 5 : training [1 2 3 4 5], test [6]

　　1，2，3，4，5，6代表的是年份。

    问2：你是怎么理解偏差方差的平衡的？

　　答：从数学的角度来看，任何模型出现的误差可以分为三个部分。以下是这三个部分：



　　偏差误差在量化平均水平之上，预测值跟实际值相差多远时有用。高偏差误差意味着我们的模型表现不太好，因为没有抓到重要的趋势。而另一方面，方差量化了在同一个观察上进行的预测是如何彼此不同的。高方差模型会过度拟合你的训练集，而在训练集以外的数据上表现很差。

　　问3：给你一个有1000列和1百万行的训练数据集，这个数据集是基于分类问题的。经理要求你来降低该数据集的维度以减少模型计算时间，但你的机器内存有限。你会怎么做？（你可以自由做各种实际操作假设。）

　　答：你的面试官应该非常了解很难在有限的内存上处理高维的数据。以下是你可以使用的处理方法：

　　1.由于我们的RAM很小，首先要关闭机器上正在运行的其他程序，包括网页浏览器等，以确保大部分内存可以使用。

　　2.我们可以随机采样数据集。这意味着，我们可以创建一个较小的数据集，比如有1000个变量和30万行，然后做计算。

　　3.为了降低维度，我们可以把数值变量和分类变量分开，同时删掉相关联的变量。对于数值变量，我们将使用相关性分析；对于分类变量，我们可以用卡方检验。

　　4.另外，我们还可以使用PCA（主成分分析），并挑选可以解释在数据集中有最大偏差的成分。

　　5.利用在线学习算法，如VowpalWabbit（在Python中可用）是一个不错的选择。

　　6.利用Stochastic GradientDescent（随机梯度下降法）建立线性模型也很有帮助。

　　7.我们也可以用我们对业务的理解来估计各预测变量对响应变量的影响的大小。但是，这是一个主观的方法，如果没有找出有用的预测变量可能会导致信息的显著丢失。

    问4：全球平均温度的上升导致世界各地的海盗数量减少。这是否意味着海盗的数量减少引起气候变化？

　　答：不能够这样说。这是一个“因果关系和相关性”的经典案例。全球平均温度和海盗数量之间有可能有相关性，但基于这些信息，我们不能说因为全球平均气温的上升而导致了海盗的消失。我们不能断定海盗的数量减少是引起气候变化的原因，因为可能有其他因素（潜伏或混杂因素）影响了这一现象。

　　问5：给你一个数据集，这个数据集有缺失值，且这些缺失值分布在离中值有1个标准偏差的范围内。百分之多少的数据不会受到影响？为什么？

　　答：约有32%的数据将不受缺失值的影响。因为，由于数据分布在中位数附近，让我们先假设这是一个正态分布。我们知道，在一个正态分布中，约有68%的数据位于跟平均数（或众数、中位数）1个标准差范围内，那么剩下的约32%的数据是不受影响的。因此，约有32%的数据将不受缺失值的影响。

    问6：你意识到你的模型受到低偏差和高方差问题的困扰。那么，应该使用哪种算法来解决问题呢？为什么？

　　答：可以使用bagging算法（如随机森林）。因为，低偏差意味着模型的预测值接近实际值，换句话说，该模型有足够的灵活性，以模仿训练数据的分布。这样貌似很好，但是别忘了，一个灵活的模型没有泛化能力，意味着当这个模型用在对一个未曾见过的数据集进行测试的时候，它会令人很失望。在这种情况下，我们可以使用bagging算法（如随机森林），以解决高方差问题。bagging算法把数据集分成重复随机取样形成的子集。然后，这些样本利用单个学习算法生成一组模型。接着，利用投票（分类）或平均（回归）把模型预测结合在一起。另外，为了应对大方差，我们可以：

　　1.使用正则化技术，惩罚更高的模型系数，从而降低了模型的复杂性。

　　2.使用可变重要性图表中的前n个特征。可以用于当一个算法在数据集中的所有变量里很难寻找到有意义信号的时候。

　　问7：协方差和相关性有什么区别？

　　答：相关性是协方差的标准化格式。协方差本身很难做比较。例如：如果我们计算工资（$）和年龄（岁）的协方差，因为这两个变量有不同的度量，所以我们会得到不能做比较的不同的协方差。为了解决这个问题，我们计算相关性来得到一个介于-1和1之间的值，就可以忽略它们各自不同的度量。

　　问8：真阳性率和召回有什么关系？写出方程式。

　　答：真阳性率=召回。它们有相同的公式（TP / TP + FN）。

　　问9：Gradient boosting算法（GBM）和随机森林都是基于树的算法，它们有什么区别？

　　答：最根本的区别是，随机森林算法使用bagging技术做出预测；而GBM是采用boosting技术做预测的。在bagging技术中，数据集用随机采样的方法被划分成n个样本。然后，使用单一的学习算法，在所有样本上建模。接着利用投票或者求平均来组合所得到的预测。bagging是平行进行的，而boosting是在第一轮的预测之后，算法将分类出错的预测加高权重，使得它们可以在后续一轮中得到校正。这种给予分类出错的预测高权重的顺序过程持续进行，一直到达到停止标准为止。随机森林通过减少方差（主要方式）提高模型的精度。生成树之间是不相关的，以把方差的减少最大化。在另一方面，GBM提高了精度，同时减少了模型的偏差和方差。

    问10：你认为把分类变量当成连续型变量会更得到一个更好的预测模型吗？

　　答：为了得到更好的预测，只有在分类变量在本质上是有序的情况下才可以被当做连续型变量来处理。

    问11：“买了这个的客户，也买了......”亚马逊的建议是哪种算法的结果？

　　答：这种推荐引擎的基本想法来自于协同过滤。协同过滤算法考虑用于推荐项目的“用户行为”。它们利用的是其他用户的购买行为和针对商品的交易历史记录、评分、选择和购买信息。针对商品的其他用户的行为和偏好用来推荐项目（商品）给新用户。在这种情况下，项目（商品）的特征是未知的。

　　问12：在k-means或kNN，我们是用欧氏距离来计算最近的邻居之间的距离。为什么不用曼哈顿距离？

　　答：我们不用曼哈顿距离，因为它只计算水平或垂直距离，有维度的限制。另一方面，欧氏距离可用于任何空间的距离计算问题。因为，数据点可以存在于任何空间，欧氏距离是更可行的选择。例如：想象一下国际象棋棋盘，象或车所做的移动是由曼哈顿距离计算的，因为它们是在各自的水平和垂直方向做的运动。

　　问13：我知道校正R2或者F值是用来评估线性回归模型的。那用什么来评估逻辑回归模型？

　　答：我们可以使用下面的方法：

　　1.由于逻辑回归是用来预测概率的，我们可以用AUC-ROC曲线以及混淆矩阵来确定其性能。

　　2.此外，在逻辑回归中类似于校正R2的指标是AIC。AIC是对模型系数数量惩罚模型的拟合度量。因此，我们更偏爱有最小AIC的模型。

　　3.空偏差指的是只有截距项的模型预测的响应。数值越低，模型越好。残余偏差表示由添加自变量的模型预测的响应。数值越低，模型越好。

　　问14：为什么朴素贝叶斯如此“朴素”？

　　答：因为它假定所有的特征在数据集中的作用是同样重要和独立的。正如我们所知，这个假设在现实世界中是很不真实的，因此，说朴素贝叶斯真的很“朴素”。

　　问15：花了几个小时后，现在你急于建一个高精度的模型。结果，你建了5 个GBM（Gradient Boosted Models），想着boosting算法会展现“魔力”。不幸的是，没有一个模型比基准模型表现得更好。最后，你决定将这些模型结合到一起。尽管众所周知，结合模型通常精度高，但你就很不幸运。你到底错在哪里？

　　答：据我们所知，组合的学习模型是基于合并弱的学习模型来创造一个强大的学习模型的想法。但是，只有当各模型之间没有相关性的时候组合起来后才比较强大。由于我们已经试了5个GBM也没有提高精度，表明这些模型是相关的。具有相关性的模型的问题是，所有的模型提供相同的信息。例如：如果模型1把User1122归类为1，模型2和模型3很有可能会做同样的分类，即使它的实际值应该是0，因此，只有弱相关的模型结合起来才会表现更好。