

订阅DeepL Pro以编辑此演示文稿。  
访问[www.DeepL.com/pro](https://www.deepl.com/pro?cta=edit-document)，了解更多信息。

好吧，我想我必须准备好这一切。 那我们开始吧 大家早上好 Hi. 大家早上好 我去给迟到的人关门吧 这是你第一次演讲吗？ 哦，天哪，太荣幸了 好了 艾丹 欢迎来到 Okay, Aidan, welcome to the. 爱丁堡大学商学院 我想我应该算是你的欢迎委员会吧 不过上周你确实有一个欢迎周 我想是的 所以 你们对学校有了一定的了解，知道了我们是怎么做事的，也许还认识了一些人。 但你们可能并不了解彼此，也不了解我，更不知道会发生什么。 所以，这对我们所有人来说都非常激动，包括我自己。 欢迎来到预测分析和数据建模课程。 今天将是一个轻松的讲座。 因此，我认为有必要先确定一下我们的期望，我们想从这门课程中学到什么、得到什么，或许还可以相互了解一下，谈谈我们在这门课程中所要涉及的内容，并介绍一下什么是预测分析。 我们将在这门课程中学到什么？ 是的，所有这些有趣的东西。 让我们先来整理一下内务。 这是课程介绍中的官方课程概述，虽然略显枯燥，但对你很有帮助。 本课程旨在培训预测分析领域的学生，使他们能够利用各种方法应对就业市场。 因此，学生的学习旅程将是通过预测分析的视角，从众多噪音中分辨出真正信号的探索之旅。 更具体地说，本课程涵盖了预测工作的典型方法步骤、统计建模、人工智能方法，还涵盖了预测分析中的实际问题以及如何解决这些问题。 因此，这种理论方法与实际问题之间的平衡将极为重要。 这也是我们将讲座和计算机实验室相结合的原因之一。 那么，这在实践中意味着什么呢？ 你们将有 20 个课时，即 11 周内有 10 次讲座，每次 2 个小时，以及 10 个辅导课时。 其中还有一周的阅读课。 我想是第六周。 也就是说，那一周没有讲座，也没有计算机实验室。 在这一周里，你们可以补上之前落下的阅读课。 第五周我们还会有一个客座讲座，这样你们就可以稍微休息一下，不用整天站在这里跟你们说话了。 我们将邀请一位大学校友，谈谈他在实际应用本课程所学到的一些技巧时的经验。 我认为，这将会给你带来一些对未来的美好展望，你可能会用在这门课程中学到的知识来做什么，也会给你带来一些对未来的希望。 因此，理论概念将在讲座中介绍，然后我们将在计算机实验室中实施。 这是一种综合方法，因此我们会兼顾理论和实际意义。 你还会注意到，讲座的应用性非常强，所以我们会讨论很多机器学习方法背后的数学知识，但我们会从实践的角度来讨论。 因此，我们不仅会讨论一种方法是如何工作的，而且还会讨论这种方法的应用意味着什么？ 这对你意味着什么？ 例如，如何为特定的数据集选择特定的方法。 因此，这里不是信息学院。 你不是在攻读计算机科学学位。 你们攻读的是商业分析方面的应用型学位。 因此，在应用分析技术解决商业问题这一特定领域，我们有两本推荐读物。 其中一本我随身带着，因为我有一本纸质版的，所以这是大部分讲座的核心内容。 这是一本相当不错的书。 这本书的一个缺点是，它使用的都是实现实例。 因此，每一章都有一个理论概述，并附有大量案例研究和实施示例，但它也使用了艺术手法来展示如何实际实施这些方法。 实践。 在本课程中，我们将实际使用 Python。 所以这也是这本书的一个缺点。 尽管如此，我坚信，如果你会一种编程语言，你可能也能读懂其他语言。 因此，即使你从未见过我们的语法，也能读懂整个语法，因为它与 Python 非常相似。 现在，我们要使用的另一本书是一本真正使用 Python 作为应用语言的书。 我也有一个关于这本书的问题，因为在你的图书馆资源中，它仍然列出了你的版本。 丢失的原因是去年这门课程确实使用了我们的资产代码语言。 所以这可能是我的一个失误。 当我们复制去年的一些信息时，我们并没有改变这本书。 尽管如此，这本书在我们的版本和 Python 版本中都存在。 它们是同一本书，唯一不同的是语法和实现方式。 因此，请尽量选择 Python 版本，但如果你有 R 版本，至少从理论角度来看，这并不重要。 你仍然可以阅读这些章节，内容也是一样的。 说到书，考虑到你是新来的，我给你简单介绍一下如何找书。 我不知道你是否真正涉及过这个问题。 有没有给你介绍过图书馆系统？ 没有。 发现这一页有点。 好吧，我们这里有两个图书馆，你们可以使用。 我的意思是，我们有更多，但两个主要的图书馆。 我们有校园图书馆，也就是主图书馆，在两栋楼的下面。 然后我们还有自己的商学院图书馆，即研究生专用的中心图书馆。 你可以在地下一层找到它。 因此，我们有自己的图书管理员，自己的图书馆系统，但我们也可以访问整个大学的图书馆系统。 这是他们的主网站。 因此，如果我们想借阅这本书，它就会告诉你在哪里可以找到。 它会告诉你在哪里可以找到它。 确实如此。 所以你可以看到这里写着在线访问和全文可用。 这意味着可能在某个地方有纸质版。 可能更多的是在国王大厦，也就是我们在城南的 Stem 校区。 不过，你可以在网上查阅这些书的全文，也可以在网上查阅这两本书的全文。 因此，如果你不想买纸质版，就没必要买。 甚至不需要从图书馆借纸质版。 在线版本完全没问题。 如果你用学生账户登录系统，就可以访问这些在线版本。 因此无需购买昂贵的教科书。 是的，评估对你来说可能也很重要。 因此，评估将是双重的。 60%的成绩来自小组作业。 然后在年底，也就是 12 月的某个时候，你们会有一次笔试，那是一次个人考试。 小组作业将由一份报告组成，报告将记录你们使用课堂上学到的技术对提供的数据集进行分析的过程。 所以我们会把你们随机分成几个小组。 不，你们不能选择小组。 我们会把你们分好，然后你们会一起研究我提供给你们的一组数据，你们会使用你们认为最适合解决所给问题的任何技术。 因此，你们可以选择技术，然后撰写一份报告，讨论和分析你们的发现。 更多细节可能会在本周末公布。 然后，你们可以随时以小组为单位开始工作。 我们会给你们一个截止日期，但你们可以提前或延后开始。 只要你们在截止日期前提交，我并不在意，请照办。 日期将会确定。 考试将在 12 月举行。 考试时间表有望在 12 月初公布。 11 月下旬会公布。 这取决于他们需要安排多少场考试。 安排所有考试是一项庞大的工作，既要保证学生的考试地点和时间不重叠，又要给我们足够的时间来制作你们为我们编写的任何内容。 是的，我说了很多却没有真正介绍自己，所以我现在要介绍一下自己。 我的名字叫安东尼娅-基钦（Antonia Kitchen），这对你们来说可能是个很难念的名字，所以你们可以叫我安东尼娅（Antonia），如果你们愿意的话，也可以叫我吉布森博士（Dr. Gibson），如果你们能念出来的话。 如果你已经从我的口音中听出来了 那是不可能摆脱的 还有我的名字 我在德国出生长大 我来英国已经八年了，所以我在尽力模仿英国广播公司的英语，但还没到那一步。 我是这所大学的博士，我的论文是关于空间时间聚类分析的。 所以我们会有一次关于聚类分析的讲座，你会看到我对这个话题特别兴奋和高兴，并请你就这个话题提很多问题。 后来我去了美国。 我在卡内基梅隆大学待了一年，做了博士后。 我在匹兹堡超级计算中心工作，这真的非常令人兴奋。 匹兹堡超级计算中心实际上是 CMU 和匹兹堡大学的联合计算中心。 我们在那里的工作是为两所大学的任何大型计算项目提供计算资源。 因此，我们经常与物理、数学和统计学系合作，同时也与医学、心理学系合作，所有拥有大量数据并希望对其进行分析的人都会来 PSC。 现在，在美国待了一年后，我意识到美国不适合我。 尽管我的工作环境很好。 但我并不喜欢美国。 所以我决定，好吧，让我们回来。 我真的很想念爱丁堡，于是就在这里做了预测分析的讲师。 实际上，你们会在本学期看到很多元数据，因为我将同时教授数据分析原理和预测分析模块，我还将为你们做预测分析计算机实验室，但在数据分析原理计算机实验室里，我将有两个了不起的塔斯来照顾你们。 我也意识到这些名字太长了。 哦，好吧。 让我喝点东西。 现在还早 好的 我已经说了很多了 我们已经确定你们互不相识 你们也不知道自己在这里做什么 所以我们先来介绍一下我自己 然后再交给你 你们可能会对我的研究领域感兴趣。 我的研究兴趣大致属于计算社会科学领域。 我认为自己是一名计算社会科学家。 这意味着我对各种方法论、机器学习和计算统计在人类行为中的广泛应用很感兴趣。 因此，有几个领域由此产生。 我的背景实际上是定量营销，所以我仍然在消费者行为方面做了很多研究，尤其是与食品消费、旅游业以及游客在一个国家各个地区的流动有关的研究。 我对经济福利以及经济福利与人们身心健康的关系很感兴趣。 因此，我们在美国实际上是在讨论如何。 你的财务状况与你的心理健康息息相关。 因此，如果你的经济状况不佳，就会影响你的精神状况，从而影响你的工作状况，进而影响你的经济状况。 因此，这是一种递归系统。 但这一领域的很多分析实际上都是分开进行的，并没有真正将这些不同的数据集结合起来。 因此，我们正在研究一种结合不同数据集的方法，以便更全面地分析这一问题。 我几乎对本地食品系统感兴趣。 我对人们如何获得新鲜农产品以及这些农产品的价格如何感兴趣，尤其是在代表人数较少的地区。 因此，我正在与加拿大的同事合作，他们正在研究加拿大的土著居民以及他们如何获得新鲜的食物。 我还对苏格兰的农村地区感兴趣，并对空间不平等这一更为宽泛的概念以及上述所有问题感兴趣。 所以，如果你对这些感兴趣，你知道你必须在硕士课程结束时写一篇论文或毕业论文。 如果你对这方面感兴趣，请告诉我，我们可以一起讨论。 也许这就是你的一篇不错的毕业论文。 现在轮到你们了。 因为你们还不认识，所以你们可能会和认识的人坐在一起。 我知道这很自然，但你们可能并不认识身边的每一个人。 所以，我希望你们转过身去，组成一个 5 人左右的小组。 我想这是一个很自然的小组。 互相交谈，如果还不认识，就自我介绍一下，然后谈谈你们想从这门课程中得到什么。 现在，这对我来说也是一个相当重要的问题，因为我想让这门课程对你们来说既有趣又适用。 所以，在你和周围的人聊了五到十分钟并讨论了你想从这门课程中得到什么之后，也许可以谈谈你的经历，你为什么选择这个课程，你为什么选择爱丁堡，以前有什么工作经历或本科经历？ 然后，我们会在房间里做一个简短的交流，挑选几个与我有眼神交流的人，然后你们可以告诉我讨论中出现了哪些话题。 不必是你自己的话题。 这就是为什么你们在小组中讨论的话题可以是你们小组中出现的话题。 如果我发现了一些东西，就有可能让我在课程中更倾向于你们感兴趣的内容。 比如，我可以把重点放在你们感兴趣的应用领域，这样你们就会更喜欢。 所以，好好讨论十分钟，然后我们再分享。 好吗？ 另外，让我再次开始录音。 我不知道你们是否知道，讲座可能会被录下来。 因此，如果你在讲座中错过了什么，这些内容会在讲座结束后上传到学习网站上，然后你就可以重新观看讲座，弥补错过的内容。 计算机实验室不进行录制，所以你应该到场。 在某些计算机实验室，还将进行考勤。 因此，我们实际上是在检查人们是否在参与，这只是参与监测。 所以我们知道，如果出了什么问题，有人突然从地球上消失了，我们是知道的。 所以，不要觉得我们对你检查得太多。 我们只是想和你们保持联系，让你们参与进来。 所以，也许从左边开始太明显了。 让我们从右边开始。 那就从后面开始吧。 你们讨论了什么？ 有机会自我介绍一下吗？ 有什么有趣的事吗？ 是的，我们讨论了很多事情。 很好 - 是啊 - That's good. 很好 我们分享了对这门课程的期望。 其中之一是，我们希望能够使用无监督或有监督的学习方法，就像我们讨论过的那样，这门课程与数据科学的步骤类似，都是从领域知识、建模和算法比较开始。 这就是我们在《完美》中讨论过的内容。 在这个过程中，你有什么特别期待的吗？ 我还没决定。 哦，天哪。 哦，天哪。 现在，这是很好的。 这是很好的。 我们将有时间。 我们将有11个星期来弄清楚这一点。 所以，这是很好的。 时间很长 我知道你不会觉得时间很长 如果你真的到了12月 课程就接近尾声了，也许就在前面。 是的。 我们讨论了在本科学习期间学到的一些模型，我们学到了一些像bagging或决策树、随机森林等，都暴露了一些问题，比如，有很多模型。 但问题是，我们不知道如何用这些模型来预测现实中的事情，我们想知道如何获取参数来预测未来，预测我们想在实验室里得到的东西。 好的，所以你知道这些模型的理论，但你想知道更多的实际意义以及如何实际使用它们。 这真是个好消息，因为这正是我们要做的。 我想决策树将在第九周进行讲解。 所以，期待第九周的到来。 你是同一组的一员，还是单独一组？ 独立小组。 所以，另一点，请。 好的 - 好的 - Yeah. Yeah. 你们中的任何一个。 Thank you. 是的，我们已经讨论过一些事情了。 是啊，我的一个朋友在这里，呃，她是在一个初创公司，他们到。 觉得这是体育之类的东西。 所以通过学习 这是分析和。 什么是预测模型？ 所以，我们可以让仪表盘不仅仅显示当前情况，还可以显示一些漂亮的或预测的东西。 还有 哦，这样啊。 这是一件事 就我而言 我也想探索一下。 我们的工作 有很多数据 我在一家电信公司工作，所以我们有很多 流量和流量数据。 所以我想知道我们如何使用这些数据，并从中获得洞察力。 这就是有趣的地方。 所以你们有实时数据流进来。 我不知道 我写下来吧 I don't know. 让我写下来。 这是一种非常有趣的数据类型。 因此，我们将简要谈谈与之密切相关的时间序列数据。 但实际上，在这种情况下思考在线学习可能也是一个有趣的点。 让我记一下。 太好了 好的 我的左边 啊哈 惊喜不同 这次 好吧 我们聊了很多 也有一些共同点 我觉得对我们来说 如果我们想知道如何预测事物 对我来说，我想，你知道，如今我们面对的是大量的数据，大量的，你知道，这些统计数据。 所以，我想知道如何从数据中抽象出有用的信息。 现在，数据只是数据，而建模只是一种工具。 所以最重要的是从数据中提取有用的信息。 我想知道具体步骤。 你是怎么做的？ 是的，我完全同意。 是的，我完全同意。 因此，收集大量数据其实相对简单，很多公司都在这么做。 所以他们只是在收集数据，因为有人告诉他们数据真的很重要。 数据是未来的趋势，但他们并没有考虑如何利用这些数据。 因此，他们很少有针对性地收集数据。 因此，要知道你究竟想收集什么样的信息，以回答你的具体问题，可能会非常困难，因为谁知道五年后你会问自己什么样的问题。 因此，考虑未来并决定未来要回答什么样的问题是相当棘手的。 因此，我想我们会在整个课程中稍微触及这个问题。 我们会讨论什么样的数据适合不同的模型，以及如何解释不同模型得出的结果。 这些也将成为你课程作业的一部分。 所以我也很期待。 左后卫？ 左后卫？ 你们中的任何一个。 我不认识你们 I don't really. 我不认识你们 你们谁都可以说 我们在谈论的是，我们非常期待，期待着这一切的应用，因为我们都很新，像Python和。 嗯，其中的编码部分。 所以，这对我们学习如何应用和从数据中得出结论是很有帮助的，因为在这方面，我们的效率其实并不高。 你明白我的意思吗？ 你明白我的意思吗？ Yeah. 好极了 这正是我们在计算机研讨会上要做的。 我想这对你来说会很有趣，我们一步一步地从如何从数据中推导到实际执行这些模型，然后解释结果，你通过同组第二。 你想要什么？ 什么？ 那个。 那种组。 好的，我们后面还有一个小组。 我想我们很多人都有同样的想法 我们都来自 外国背景 呃，默认情况下可能是计算机科学。 一般来说，分析对我们来说可能比较难，但我们确实很期待。 呃，不管我们是用数据来创造瘾还是。 然后，一般来说，在哪些应用中使用适当的方法。 我们也期待着使用现实生活中的例子，尤其是考虑到很多场景都涉及到几个异常值，我们应该如何处理以及适当的处理方法。 是的，是的。 我想，异常值会在第三周作为一个重点出现。 但在整个系列讲座中，我们都会谈到异常值。 你们都提到了不同的背景，这真的很有趣，让我看看谁更有计算机科学数据科学统计学的背景。 让我们把所有的 Stem 都称为 "Stem "吧。 好的，一半左右。 那谁的社会科学业务更多一些呢？ 背景类似，另一半。 所以，你可以看到，我们在这门课程中实际上要做的是把双方结合起来。 所以你有时会想，哦，这对我来说更难，这对我来说更容易，这完全正常。 而对你的同事来说，情况恰恰相反。 因此，要试着在这两类人之间进行混合，试着与你的同事交流，尤其是那些背景不同的同事。 如果你觉得讲座理论家中的某个部分特别具有挑战性，其他人会比你更容易理解，他们也能向你解释。 你们可以一起学习。 我认为，这才是真正的系列讲座方式，将其视为这两个领域的结合。 好吧，在我没听清楚之前，你是一个单独的小组，还是属于背景小组、前台小组。 你属于这个群体。 所以我涵盖了所有人。 好吧，那么中间，中间，后面的地方。 好的 所以其实12个人也在一组。 你知道，其实最后一个的问题是 我想每个人都已经说过了。 我是说 这对我来说挺好的 不过 实际上 我们五个人都来自不同的背景 就像你说的 多样性 我们中有些人有计算科学背景 其他人则是最大值 我们中的一些人，你知道的，也有一些人主要熟悉理论方面的东西，并期待着实践部分。 另一种情况是，我们中的一些人已经从数据本身入手，开始了解理论方面的知识。 还有一件事。 就我而言，我来自理论界，这意味着我总是能得到一套现成的数据，我只需用简单的程序进行分析、预测即可。 但现实生活中的数据显然不是这样的。 是啊。 所以我很想知道如何清理数据，如何让数据随时可用。 最后，还有一点是我之前说过的，我们在讨论这门课程之后，希望能把业务方面的想法都带进来，因为这是这样的，因为碰撞总是这样发生的，业务人员是怎么想的，分析人员又是怎么想的。 那么知识实际上是不匹配的。 但希望在这门课程之后，我们能将两者结合起来。 大家都很高兴。 这只是知识和语言的问题。 因此，在很多情况下，你实际上使用完全不同的语言来表达相同的概念，这真的非常非常棘手。 所以有时你会意识到你们是在并行对话。 所以，你们用不同的词来谈论同一个概念，你们互相生气，因为看起来你们不理解，但实际上你们在谈论同样的事情。 这很令人沮丧。 因此，我们将尝试弥合这种差距，也许会给你一些词语和工具，以便能够与双方交谈，是的。 后面，右边 我的右边，你的左边 好的 所以，首先，我们讨论的权力。 我们很难在这种寒冷潮湿的天气里生存 因为你来自哪里？ 我来自库尔德斯坦 那是中部地区 我想我们组的人也是来自同一个方向，来自同一个方向。 所以，我是说，喜欢。 此外，我们还讨论了比较艺术编程和飞行时间。 我们。 关于语言。 是的，我们讨论过 Python 在使用上要简单得多，但也有很多优秀的研究论文。 在我们的编程中。 所以我们只是都知道在实现。 到模型，我们已经做了。 你知道如何教我们如何实现。 冷得要命 哦，我的天哪。 太好了 我不知道你听说了什么 今年夏天真的很热 气温高达26度 对苏格兰来说热得难以置信 苏格兰真的非常非常热 另外，你可以试试住在挪威。 那很有趣 我曾经 我在零下27度的时候还在讲课，太冷了 好吧 你说的不是零下20度 Okay, so you're not talking -20. 我们现在说的不是零下十度 We're not we're not talking about ten degrees now. 是的 我感觉到了 Yes. 我感觉到了 适应新的气候很困难 所以，如果你们喜欢，就穿上外套吧 So keep your jackets on if you like. 没关系 我把后面也盖上了 我想是的 那么 你们在前面是一组吗？ 好吧 给你点东西 Okay. 给你的东西。 我们主要谈了两个方面。 一个是，我们大多数人都在本科期间或工作经历中实施过一些其他模式。 但我们 真正想了解的是 一个特定的 模式是否。 是正确的 还是不对。 所以早在我工作的时候，就有一些人刚。 左，右。 实施模型，但他们。 并不真正理解背后的数学。 所以这。 有点 奇怪，因为后来我开始学习的时候。 所以，这是我们要学习的一件事。 第二件事是了解特定模型的实施对业务的影响。 因为，你知道，你只需向你的客户推荐一种模式，然后说，嘿，这真的很花哨，这一切都很有效。 但客户必须了解其重要性，以及是否与他们的业务战略相关。 因此，这一点变得非常重要。 所以，这就是我们作为一个小组期待从这门课程中学到的两件事。 非常好 是的，你说得非常好。 我觉得现在每个人都想学习神经网络和深度学习，所以每个人都会告诉我，我们什么时候开始讲深度学习？ 然后你就会看到这些公司带着他们的小数据集，想要对非常简单的关系进行深度学习。 这完全没有意义。 所以，我们实际上会介绍什么时候使用什么样的模型才有意义？ 有时，一个简单的模型确实更有意义，因为它的结果、准确性往往更好，也更容易解释。 它适用于较小的数据集。 因此，在很多情况下，简单的模型更有意义。 在某些情况下，你需要深度神经网络。 如果你要分析具有非常复杂的非线性关系的大量数据，深度神经网络就非常有用。 但我觉得这是我个人的一个小毛病。 我觉得它们真的被过度使用了。 所以，在整个系列讲座中，你不会再听到我对神经网络津津乐道。 希望这不会让你们失望。 好的，很高兴听到大家的意见。 我认为，能听到同事们的意见也非常有价值，因为你们都有相似的思维方式，都在同一条船上。 那么 让我们来谈谈今天讲座的其他内容。 我们将讨论什么是预测建模，以及本课程将涉及哪些相关内容。 我们将简要介绍不同类型的模型，以及整个建模过程的结构。 为此推荐阅读前几本书。 这两本书的前两章，如果你想看看的话。 现在我们来谈谈预测建模。 我们已经说过，预测建模实际上是一种广义上的决策制定，因此可以涉及到公司内部的决策制定，但也涉及到你日常生活中的决策制定。 因此，每天你都会在这里做出一些决定，比如我应该走哪条路去校园，我也不知道，最快的是哪条路？ 我今天应该穿夹克还是带雨衣或雨伞去苏格兰？ 答案总是肯定的。 但为什么总是肯定的呢？ 因为根据历史、根据天气预报、根据以往的经验、根据你对气温的个人感受。 现在，所有这些都在影响着你在日常生活中如何做出决定。 那么，我们如何做出决定呢？ 我们收集信息，这些信息可以来自不同的渠道。 例如，如果我们决定如何从 A 地前往 B 地，我们可能会询问谷歌地图，而谷歌地图可能会告诉我们：嘿，这条路目前关闭了，所以今天换一条路线吧。 因此，根据谷歌地图提供的信息，我们会改变决定。 同样，如果我们想在当地一家面包店买到最新鲜的面包，我们可能会查看 Facebook 上的评论，这些评论会告诉我们：嘿，他们总是在上午 11 点制作面包。 或者你会想，我下一步该买什么样的车？ 然后你可能会问父母，可能会问母亲，因为他们会向你推荐某一款车。 然后根据她的经验，你可能会检查一些你知道自己想要的新车的技术指标。 然后在此基础上，你就会决定购买特定的汽车型号。 所以。 所有这些情况都会用到信息和数据，但它们的来源各不相同。 有些数据是客观的。 有些数据是主观的，有些数据是两者的混合。 因此，举例来说，征求推荐意见可以获得主观数据。 这只是一种意见。 对你来说，这可能是一个非常重要的意见，你可能会非常信任这种数据和意见来做出决定。 但与技术规范相比，这仍然只是一种意见。 这些都是客观数据，因此它们客观上反映了汽车的真实情况。 至于如何解读这些客观信息，还是取决于你自己。 在这种情况下，你可以看看它是否符合你的个人喜好。 例如，这需要什么？ 你必须对这些规格有一定的了解。 因此，为了真正解读数据，你需要知道数据的来源。 你需要知道数据的可靠性有多高，数据的相关性有多大。 你还需要知道如何解读数据。 因此，这实际上是一个相当困难的决策过程，而这些都是你每天都要做的日常决策，也是很简单的决策。 因此，如果我们考虑一下比这更难的决策，在某些时候，你会发现你作为一个人无法独立回答这个问题。 这里有几个问题。 例如，我们应该为即将推出的新产品投放什么样的广告，或者我们应该投资这只特定的股票，又或者五年后的房价会是什么样？ 我想我们都希望房价会比现在低一点，因为我现在还贷不起。 因此，所有这些问题都非常棘手。 而我，作为一个普通人，也许可以做一个有根据的猜测，但我真的不能可靠地告诉你，某家公司的股票在几个月后会是什么样子。 为此，我会尽可能多地收集有关该特定问题的数据，然后使用预测模型，根据这些我自己无法分析的数据量来预测未来。 这就是预测建模背后的整个理念。 我们正在改进你大脑中已有的决策过程，使其适用于更广泛、更复杂的问题。 不过，我之前提出的问题依然存在。 信息来自哪里？ 有多可靠？ 相关性如何？ 如何解释结果？ 你需要什么样的知识才能解读结果？ 所以，这些都是一样的，都是正确的。 我的书，我最信任的书是这么说的： 预测建模是开发一种数学工具或模型来生成准确预测的过程。 准确性是我经常提到的关键词之一，因为我们实际上是通过它来比较大多数模型的性能的。 因此，我们要看模型预测的结果与实际情况有多接近。 尽管如此，有时更高的准确率并不是你应该一直追求的最佳目标。 在机器学习中，你经常会看到一种猎取准确率的行为，所以你可能会看到几篇论文，它们报告的准确率分别是 78.3% 和 78.6%。 显然，第二个模型更好，我们应该始终选择第二个模型。 在这种情况下，人们并没有真正想到的是，那个模型有多昂贵？ 该模型需要什么样的数据才能执行？ 这就是黑箱模型。 监管机构不喜欢我们在金融领域使用黑箱模型。 因此，所有这些问题也非常重要。 这并不全是准确性的问题。 在评估模型时要综合考虑各种因素。 这就是理想化的预测建模过程，你会以某种形式大致看到这个过程。 在很多书中都可以看到。 我们从某种问题开始。 我们收集与问题相关的数据。 我们对数据进行预处理，查找异常值，进行归一化处理，等等等等。 选择模型、训练模型、评估模型并报告结果。 现在，实际情况更像是这样，更像是循环，因为当你真正开始报告结果时，会出现更多问题。 很少有人能最终找到你一直在寻找的答案，解决所有问题。 新的问题又会出现，整个过程又会重新开始，甚至可能比这更早开始，因为你意识到没有数据可以真正回答问题，或者数据质量很差。 因此，你必须重新开始收集数据，这就是整个循环过程，在这个过程中，你会问自己很多问题。 那么，如何将问题转化为问题，这又回到了我们之前谈到的语言问题？ 商业问题的用语可能相当模糊，而试图将其表述为一个可测试的问题是非常具有挑战性的。 数据收集 哪些数据是相关的？ 如何找到这些数据？ 我能使用这些数据吗？ 对这些数据有哪些伦理方面的考虑和问题？ 然后，你会问自己：我还需要更多数据吗？ 然后你又会问自己：我的数据是不是太多了？ 因为笔记本电脑经常死机？ 所以，到了某个时候，你就会忘了自己真正想问的问题，转而思考我为什么要这么做？ 我走的路还对吗？ 或者到底发生了什么？ 所以你不得不继续前进，因为这就是你要做的。 你会想，我是否应该选择一个完全不同的模式？ 这种模式适合这个问题吗？ 最后，你会问自己，这一切意味着什么？ 我该如何包装？ 我为什么要这样做？ 我为什么来读研究生？ 诸如此类。 然后你又重新开始。 这就是我过去六年来一直在做的事情。 好了，以上几点我都说过了，从业务问题开始，决定你需要的数据，收集这些数据，预处理数据，选择模型。 这实际上也很棘手，因为模型的选择取决于很多不同的因素，而在很多情况下，这些因素都是主观和非理性的。 与其他模型相比，有些人更喜欢某种类型的模型。 我之前提到过，我非常喜欢聚类和无监督学习。 除了我花了五年的时间来做这件事之外，没有任何合理的解释。 所以我擅长它，所以我更喜欢尽可能使用它。 这是不合理的。 训练一个模型也是相当棘手的，实际上，这也是很多人在考虑数据科学和分析过程时非常关注的一点。 他们都在考虑训练这一步，或者建立模型这一步。 他们并没有真正考虑到评估结果之前和之后的步骤。 我已经谈到了准确性问题。 报告结果也很有趣，因为你要把问题从商务英语翻译成数据科学英语，然后再把结果翻译回商务英语。 因此，这实际上是非常有趣的思考，也将是你们小组项目报告撰写的一部分。 所以，你们还是要好好想想，如何将研究结果转化为可视化的东西，或者如何与人很好地交流。 外面太吵了。 我以为我得停下来。 但我还能继续 好吧，这就是问题所在。 让我们用非常正式的术语来谈谈预测模型的组成部分。 预测模型其实就是试图预测变量之间的关系。 这就是我们要做的。 因此，我们有一些用 X 表示的特征，它们可以被称为特征、解释变量、自变量。 这就是你的数据，这就是你为解决问题而收集的数据。 然后，你就有了某种目标。 你可以称其为标签、响应、变量输出结果，随你怎么称呼。 这就是因变量，因为它取决于自变量。 很显然，你还需要某种误差项。 所以，没错，就是符号。 我们用粗体字母书写向量和矩阵，矩阵大写，粗体向量小写，粗体字母。 这只是为了放在墙上。 所以，你的幻灯片里已经有了。 有趣的是函数 f，它描述了因变量和自变量之间的关系。 这就是我们正在构建的，也是我们实际上正在尝试建模和创建的。 因此，数据与问题或问题答案之间的关系，就是我们要建模的内容。 我们通常会这样做。 先用我们的数据和自变量建立一个过程，然后再用实际存在的已有结果样本进行测试。 因此，我们有一个给定的 YS 样本，我们有一个给定的 X 样本，我们正在建立一个模型，然后我们正在测试该模型的性能如何。 如果我们给它相同来源的新数据，但没有结果，那么我们就可以将结果与新数据、旧数据的结果进行比较。 这就是预测建模非常简单的基本过程。 我们还谈到，不同领域的术语可能会有所不同，不仅是商业、社会科学领域，例如，商业专业人士可能来自不同领域，而且不同领域内的术语也会有所不同。 例如，我们正在讨论 R 和 Python 之间的区别。 更早一些。 这很大程度上取决于人们来自哪个领域，以及他们究竟来自哪个领域。 统计学家经常使用 R。 例如，数学家通常使用 Matlab，而大多数计算机科学家使用 Python。 因此，如果你想把统计学家、机器学习者和数学家放在一个房间里，让他们创建一些东西，这将需要一段时间，因为他们会不停地争论语言的使用，然后他们会谈论术语。 因此，这实际上相当棘手。 在我攻读博士期间，我的导师是一位统计学家和一位数学家，幸好他们相处得很好，这对我这个学生来说是件好事。 但我们有时还是会遇到这样的情况，他们不得不互相交流，互相倾诉。 好吧，矩阵的符号和我习惯的不一样。 应该遵循。 我们应该告诉学生什么？ 她应该使用哪种符号？ 因此，我们对此进行了大量讨论。 这里有几个数据点的例子。 有些人称它们为样本。 因此，一个单独的数据点实际上被称为一个样本，但你取样的多个数据点也被称为一个样本。 所以我不是很喜欢这个词。 你也可以称它们为观测值、实例或测量值。 特征被称为预测因子、自变量、输入属性、描述符。 因变量可称为目标类结果或响应变量。 因此，所有这些都取决于你读的是什么书。 你在读什么样的论文，你的背景是什么。 我并不特别在意你想用哪种语言，所以你可以随意选择一种，也许是你最喜欢的，或者是你更喜欢的。 我很可能会使用机器学习方面的术语，但也可能会偷偷加入一点统计统计学的内容，因为这是我的专业背景。 因此，这取决于我是在用机器学习术语还是统计术语来思考问题。 不过，如果你感到困惑，可以问我，我可以试着翻译成你的语言。 我只会说两种半语言。 真的吗？ 是的 同样，数据类型也有不同的说法。 分类数据可以称为名义数据、属性数据、离散数据、定性数据，等等等等。 有时也有一些特殊的数据类型，比如李克特（Likert）。 社会科学中经常使用量表。 它们只是顺序变量的一个子类型，但由于它们用得太多，你有时会看到术语 "李克特量表数据 "泛指顺序数据，尽管它指的是一种非常特殊的七分量表。 例如，模型训练、建立参数估计，所有这些都是指同一个过程，严格来说，参数估计是一个子过程，但人们用它来指代一般术语。 好了，刚才说了很多术语，我们继续。 我们将讨论有监督无监督建模，我想我们之前已经略有涉及。 有监督建模和无监督建模是机器学习的两大类，严格来说，还有半监督学习，以及其他一些子领域。 但我们还是把重点放在这两大类上，因为这也是我们在本课程中将会涉及的监督建模。 我前面描述的过程实际上就是正在发生的过程。 我们收集了一些观测数据。 因此，我们收集了一个数据集，并根据目标变量为数据集的一部分贴上标签。 然后，我们拟合一个模型，在已知数据和每个变量的目标结果之间建立关系模型。 然后，在新数据上使用这个训练有素的模型，检查它是否仍然表现良好，是否仍能为每个新变量提供正确的术语和正确的结果。 这样做的目的是，在某些时候，你会得到一些未标记的新数据，而你仍然希望能够对这些数据进行标记。 因此，一个简单的例子是，如果你有一个二元结果，即收到的邮件是垃圾邮件还是非垃圾邮件，你有一堆邮件，你知道一封邮件是垃圾邮件还是非垃圾邮件。 因此，你可以将其放入一个模型中，然后训练该模型，使其能够根据样本识别垃圾邮件。 但随后又有新邮件进来，你又想用同样的模型来预测这封邮件是不是垃圾邮件。 根据你的邮件样本。 你已经发现了其中的问题。 根据你的垃圾邮件或非垃圾邮件样本的好坏，他们也许能更好地预测新邮件的到来。 而且您可能还会发现，5 年或 10 年后，垃圾邮件可能会完全不同。 因此，根据新数据不断更新模型非常重要，而且成本可能相当高，这取决于数据环境变化的速度。 无监督建模是一种不同的过程，因此我们没有响应变量。 因此，我们不会将数据定位或标记为垃圾邮件。 例如，我们更感兴趣的是对数据中存在的模式进行探索性分析。 例如，我们试图找到数据点的聚类，这些数据点彼此相似，但与其他数据点非常不同。 一个常见的例子就是客户细分。 基本上，你可以查看客户和他们的人口统计数据，然后将这些数据输入一个无监督模型，例如聚类算法，该算法将在数据集中寻找模式，以及彼此相似而与其他人截然不同的人群。 现在，我们并没有要预测的目标。 我们试图探索和发现数据中的群体和模式，而不需要第二步。 这样做的好处很明显，我们不需要给任何数据贴标签，所以你不需要知道某些东西是垃圾邮件还是非垃圾邮件来训练你的模型，因为我们只是在探索。 缺点很明显，你并不是在训练一个模型来预测特定的结果变量。 这更多地是为了了解数据及其结构。 因此，在本课程中，我们将介绍有监督和无监督模型。 有监督模型将包括回归 K 和 NS 决策树、支持向量机和神经网络，无监督学习将包括第一领域的聚类分析。 我们还将介绍 PCA。 是否属于无监督机器学习取决于你对机器学习的定义，因为它是一种降维技术。 严格来说，它不是一种学习方法，而是一种降维方法。 但有些人把它算作无监督机器学习。 所以我们还是相信他们的定义吧。 好的 几点了？ 要不要休息一下喝点水？ 我们做得怎么样？ 累了吗？ 对不对？ - 对 - Right? 是啊 - 好吧 - Yeah. 我也是 - Okay. 我也累了 我们休息五分钟 然后再回来看几个例子 谢谢 Thank you. 好的 好吧，让我们再次开始。 安静 安顿下来。 别担心 Don't worry. 你几乎已经完成了。 It'll be fine. 好了 准备好了吗? Okay. 准备好了吗 Ready? 好了吗？ - 好了 好了，我们刚才简单讨论了有监督学习和无监督学习，以及两者之间的区别。 让我们来看一个简单的例子，也许会更清楚一些。 这是一个有监督问题的例子。 我们正试图预测一名游客参观苏格兰某个旅游景点的可能性。 我们收集了以下数据，这些数据非常小，但确实存在。 我们这里有一个小数据集，五条记录。 我们有游客的年龄和国籍，这是我们的标签。 因此，我们实际上是在给每条记录贴标签，标明他们是否去过该景点。 现在，模型正试图找出 X（即年龄和国籍）与 Y（即是否去过该景点）之间的关系。 那么，谁想猜一猜？ 这个模型可能会寻找什么样的规则？ 你有没有发现游客和非游客之间的行为模式？ 还有年龄。 年龄 年龄 很好 你觉得H. H. H. 就像 你能再说一遍吗？ 是不太可能访问的地方。 所以所有的人都不太可能去 是啊。 举个例子，我们这里有两个年龄最大的人，他们都不去旅游，我想，是的。 年龄和国籍 英国人不太可能去苏格兰人去的地方。 是啊。 所以我们这里有两个英国人，他们也都不去旅游，而我们这里有威尔士人和苏格兰人，他们都去旅游。 这其实很有趣，因为我们可以看到这两个人都指向同一个方向。 现在，让我们想象一下，我们收集了某人的新数据。 我们有两个新的人，一个是 70 岁来自苏格兰的人，另一个是 20 岁来自英格兰的人。 你认为模型会如何预测这两个人？ 他们会去还是不去？ 会 是啊 所以你认为国籍比H更重要？ 是的 合理的猜测。 还有其他猜测吗？ 这真的很难说，不是吗？ 你并不知道是什么在驱使你的行为，是有人在你身边，还是有人来访。 那么，到底是哪个变量驱动了他们的行为呢？ 所以，这其实是完全合理的猜测，特德 可能是国籍，可能是年龄。 从这些数据中我们真的无法判断。 因此，这是一个很好的例子，说明数据的局限性限制了你的模型能够从哪些关系中学习。 在某些情况下，模型可能会随机分配。 因此，这可能是一个模型准确性不高的例子，因为这有点像折腾，除非我们收集更多数据，观察到更强的模式。 因此，如果我们能确认其中一个人是否来过，就能为我们的模型锦上添花，改善我们未来的学习。 让我们来看看我之前提到的无监督模型。 实际上，它在细分消费者方面用得非常多。 这是银行收集的一组数据，涉及移动银行应用的使用情况和客户年龄。 如果你看一下，你觉得无监督机器学习模型可能会发现什么样的模式？ 是的，这个人不太可能使用手机应用，你知道吗？ 他们可能指的就是这种大致的线性关系。 关于无监督学习，我之前是怎么说的？ 我们在寻找什么？ 模式和群体。 因此，这实际上是一个很好的例子，说明无监督机器学习模型可能会告诉你一些你并不感兴趣的东西。 因为我认为，例如聚类，会对这些数据做类似这样的处理。 它可能会找到三个组：这些点、这些点和这些点，因为它们最接近。 实际上，它并不能回答这类问题。 我们可能会问两个变量之间的关系。 它只是在行为相似的人群中找到模式。 现在，你可以对这些群体做的事情是，你可以看看蓝色群体，你会发现他们更年轻，更多使用应用程序，你可以看看红色群体，你会发现他们更年长，他们没有那么多使用该群体。 但这只是解释结果的第二步。 因此，你得到的实际建模结果只是对数据进行了纯粹的细分。 实际上，它并没有直接捕捉到线性关系。 举例来说，如果你使用一个监督学习模型，它或许能真正捕捉到两个数据点之间的关系，尤其是如果你试图在其中穿插一条线性回归线的话。 我希望大家注意的另一件事是这里的两个人，因为他们是离群值。 他们都是非常奇怪的人。 只有一两个人的行为与其他组员不同。 他们没有跟随趋势。 因此，公司现在面临的问题可能是，这些人在做什么？ 他们为什么和其他人不一样？ 是收集错误，还是他们实际上是一个有效的额外群体，我们可能会从银行的角度对他们的行为感兴趣？ 例如，在许多情况下，银行会尝试针对其目标受众或客户群体中的子群体提供特定的产品和通信服务。 例如，他们可能会为老年人创建某种沟通方式。 嘿，你试过我们的移动应用程序吗？ 这真的很好用，他们可能会把目标锁定在年轻人身上，例如，他们会通过移动应用程序来推送新产品，因为他们知道这样做可以接触到合适的受众。 但我们该如何应对这些人呢？ 这其实是一个商业问题，不是吗？ 我们要瞄准他们吗？ 我的意思是，我们可能会认为这是足够有价值的，或者我们可能会认为针对他们非常具体或特定的通信设备而没有真正得到多少回报是昂贵的。 因此，这就需要提出两个问题。 他们为什么在这里？ 他们是真实存在的，还是收集错误？ 这不是模型能回答的问题。 你必须结合对业务的了解来回答这个问题。 好的。 现在，让我们来谈谈区分不同类型模型的另一种方法，特别是回归模型和分类模型。 它们通常都属于监督建模的范畴。 因此，我们将更多地讨论有监督建模，而不是无监督建模。 它们的概念是相同的。 因此，我们希望根据某种已知数据集进行预测。 这就是我们今天已经讨论过的内容。 不同之处在于，在分类法中，我们试图预测一个类标签，因此数据点属于某个特定的类，而在回归法中，我们试图预测一个连续的量。 所以基本上是一个数值，一个结果。 有时会有一些重叠，你会看到这就是系列讲座的内容。 例如，我们会讲到线性回归，这是一种回归，一种回归模型，然后我们会讲到逻辑回归，这是一种回归模型，从某种意义上说，它给你提供了一个类别标签。 然后，我们谈谈回归树、决策树，预测一个数字。 我们会讨论决策树，预测一个类标签。 因此，我们总是在两种不同的语境中谈论同一个模型，而选择哪一种完全取决于你想要预测什么。 让我们来看四个例子，我问你这是回归问题还是分类问题。 第一个例子是，下个月特定公司的股票价格会是多少？ 回归？ 是的，这是一个我们试图预测的数字。 我公司的客户会流失吗？ 是的，这是分类的一种特殊情况，即二元分类。 他们要么会流失，要么不会流失。 只有这两种选择。 这部电影属于哪种类型？ 那么分类和这个博物馆会吸引多少游客呢？ 是的，这很简单。 但这些都是你要尝试回答的问题。 因此，在第一步，你会得到一个商业问题。 第二步，你要尝试根据这个业务问题提出一个可衡量的问题。 然后，你必须决定是用分类方法、回归方法还是某种细分聚类、无监督方法来解决这个问题。 这就是你需要考虑的三种选择。 让我们来谈谈变量的类型。 一般来说，我们首先要区分定量数据和定性数据。 现在，根据你的背景，你可能更熟悉其中一种类型。 定量数据是指任何数值化的东西，这些数字具有某种数值意义。 因此，我们在计算某些东西。 我们有一个比率。 我们有一个实数，一个通过数值记录下来的东西。 定性数据是非数值数据，在定性分析中，我们也将其视为非数值数据。 这就是最大的区别。 实际上，这是我前两天晚上在思考这个问题时添加到这个定义中的，因为定性数据的例子包括文本、文字记录、图像、声音记录等。 等等。 但所有这些数据都可以被视为定量数据。 例如，你可能读过关于图像识别的文章，也可能读过关于文本挖掘算法的文章。 因此，定量方法与定性数据之间的区别在于我们如何处理定性数据。 在社会科学领域，一般有两种观点。 你可以按照定性研究的方法处理定性数据。 例如，这意味着对某人进行深入访谈，然后分析访谈中出现的代码和主题。 或者，你也可以从一群人中收集大量的调查问卷，然后用各种定量方法对调查结果进行分析。 这两种方法都可以用于类似的问题。 因此，选择哪种方法取决于你想要得到什么样的答案。 这两种方法也都是。 这一点非常重要，是真正有效的研究选择。 话虽如此，本课程将侧重于定量研究。 因此，请不要试图说服我你想在小组研究中进行访谈，因为这与本课程的应用无关。 但一般来说，定性研究和定量研究在社会科学中同样有效。 这就是我在这里所说的。 如今，在图像识别中，许多定性数据可以并正在转化为定量数据。 这通常采用矩阵的形式，用 0 和 1 表示图像的像素。 因此，这是一种将图像转化为定量数据的方法。 然后，我们可以用常规技术对其进行分析。 例如，你可以使用神经网络来识别笔迹。 我们也可以对分类数据进行分析。 因此，如果有人拿着一个盒子问他的最高学历是什么，我们就可以将其解释为定性数据。 因此，我们可以将其解释为单词及其背后的含义。 或者我们可以将其作为一个标签，然后使用定量技术来分析这个标签。 是啊 最大的区别在于，我们是要更深入地分析少量信息，还是要分析大量数据，并从中进行归纳总结。 例如，如果你对客户对公司或特定产品的意见感兴趣，你可以对他们进行深入访谈，但你只能访谈这么多人。 因此，通常情况下，你可能会采访 1020 人，但要非常非常深入。 你会与他们交谈一个小时，了解他们的主观意见、动机和信念，以及这些是如何影响他们与公司的联系的。 这些都是非常有价值的信息，是你无法通过调查获得的。 另一方面，从这些信息中很难推断出普通受众。 因此，举例来说，使用文本识别技术从大量的文字记录中学习一般主题，而不是进行人工编码和思考，这样也许能获得更多信息，但也更表面化。 因此，让我们把注意力集中在定量数据上，因为这才是本系列讲座要讨论的内容。 我们通常将数据分为离散数据和连续数据，在离散数据中，我们又将其分为二进制数据。 零一分类数据，也被称为名义数据、序数数据，给你一个等级或例如一个液体标度，然后是数字整数账户数据。 这些都是可数的数字。 如果你回想一下我们之前的例子，参观我博物馆的游客，这就是可数的。 你可以计算到底有多少人参观。 连续数据要么是区间数据，因此零没有真正的意义，要么是区间数据的比率示例。 典型的例子是温度和摄氏度。 有一个零，但零并没有真正的意义，它只是另一个数字。 因此，温度可以低于零。 有真正零的比率的例子是重量。 重量为零意味着没有重量。 因此，它是不存在的。 这就是两者的主要区别。 是的，我提到在我们的分析中，分类变量通常会进行转换。 因此，我们正在研究一种热编码。 通常情况下，我们可以在建模中使用分类变量。 好的 我们就快完成了。 你做得很好。 我知道你们累了。 让我们来看看几种数据类型。 我之所以写最有可能的数据类型，是因为严格来说，它们可以用不同的方式记录，但每种数据类型都有最合理的解释。 我已经提到过以摄氏度为单位测量的温度。 是什么呢？ 是的，连续数据，更具体地说是区间数据。 是的。 那么主题公园的游客数量就是离散的。 是的。 在这个账户中，数据教育水平。 是的，这其实是个有趣的问题，因为你可以说，根据记录的教育水平，可能会有同一水平的类别，而这些类别不是序数，在这种情况下，它将是分类数据。 但在人口普查数据中，通常记录的是顺序变量。 餐厅游客样本中最喜欢的苏打水口味。 分类数据。 是的，很有可能。 所以提问的方式可能是，你更喜欢哪种苏打水？ 然后人们就会拿起一个盒子。 因此，这将是一个项目在一段时间内的分类价格。 是的，可能是连续变化，也可能是时间序列数据。 因此，这是一种特定的数据类型，你需要收集同一对象在一段时间内的数值，以及病人是否患病。 正是二进制数据。 我想简单提一下的另一件事是，在社会科学领域，很多情况下，你会在数据集中遇到混合数据类型，这确实很有挑战性。 因此，你可能会有某种分类响应变量，然后有一个数字和顺序输入的混合物，试图预测该分类响应，或者你可能有一个回归问题，或者你有一个连续输出，但你有数字和二进制输入。 老实说，如何将这些不同的数据类型结合起来，是一个全新的研究领域。 实际上，我最近花了相当多的研究时间来思考不同的数据类型，以及如何将它们组合成一个数据源。 如果你研究的是不同的数据源，这就会变得更加困难。 因此，在我最近的一个研究项目中，我们正在研究如何将不同国家以不同方式收集的人口普查数据结合起来。 我之前提到过教育的例子。 不同的国家有不同的教育水平分类方法，但如果你想将这些数据结合起来，并对不同国家的人口进行比较，就必须找到将两者联系起来的方法。 而这主要是由像我这样的穷人手工完成的，他们不得不翻阅数百页的人口普查文件，试图找出这些变量是如何被记录、编码和抽样的，然后试图了解加拿大和法国的教育体系，并试图了解可以通过哪些方式对它们进行比较。 因此，这真的非常非常棘手。 值得庆幸的是，除非你选择撰写论文，否则你很可能不会被迫这样做。 这就是为什么你只能研究如何处理不同类型的变量，进行预处理。 例如，我们在讨论如何进行热编码，然后在解释阶段，思考如何从分类变量和顺序变量中获取信息将非常有趣和重要。 举例来说，一个普通变量，除了排序之外，数字没有任何意义。 举例来说，高中文凭并不比硕士学位小多少，尽管它可能被这样编码。 它只是学历低于硕士学位。 但这两者之间并没有真正意义上的距离。 那么，顺序变量究竟能告诉我们什么，哪些模型适合混合数据类型？ 我觉得有人对决策树很感兴趣。 决策树非常适合混合数据类型。 是啊，永远、永远、永远检查你的模型假设是非常重要的。 我们之前也讨论过这个问题。 当你想让同事提及时，人们就会使用任何模型，而不考虑它是否真的适用。 很多人实际上就是这么做的。 他们认为，神经网络太神奇了，于是就把数据扔给它，而不去考虑它的影响。 举例来说，人们可能会有几个二进制变量，然后通过组合这些二进制变量来创建更多变量，并加载到模型中。 但你不能这么做，因为你会遇到多重共线性（Multicollinearity）的问题，即变量之间相互关联，你不能把它们分开处理。 因此，就会出现很多问题，人们会想，这是我想使用的模型。 这就是数据。 我只需将数据输入模型，然后解释结果，你可能会得到非常非常好的结果，但这并不意味着什么，因为你没有检查你的假设。 所以请检查你的假设，尽管这有点让人头疼。 所以你必须检查，比如，我不知道，线性回归误差项的独立性，等等等等。 有一个假设列表。 你必须一一检查。 你必须报告你是否检查了所有这些假设，然后你必须使用该模型并报告结果，因为否则人们就无法真正做到这一点。 法官。 你的分析做得如何，因为你可能做得很完整。 是的。 所以在 所以在。 混合数据类型。 具体来说，你会在所有步骤的预处理中遇到问题。 我们已经提到过了。 你究竟应该为模型选择哪些变量？ 有没有一种方法可以选择既适合模型又能提供大量信息的变量？ 您可能需要考虑专家知识等问题？ 这些变量究竟意味着什么？ 这可能意味着要询问相关人员、与公司沟通、阅读 200 页的文档以及所有模型部分。 我们谈到了决策树在处理分类数据和解释数据时的优势。 如果我们知道在建立模型时涉及了混合数据类型，我们该如何解释这些数字呢？ 好吧，最后一点。 最后一点。 你几乎完成了挑战，非常简短地谈到了挑战和预测建模。 我从书中摘录了这些问题：数据不足、预处理、模型验证不足、不合理外推以及模型与现有数据过度拟合。 这是作者在预测建模中经常看到的四个核心问题。 这些问题我也都见过，但我还想补充以下几点。 我自己的经验。 数据太少或太多都不行。 有一个甜蜜点。 我想我曾经画过一幅图，看起来有点像发生这种情况的概率，然后是数据太少、数据适量和数据太多。 所以，要么数据太多，要么数据太少。 我从来没有过恰到好处的数据。 你可能也不会遇到这种情况，所以这很棘手。 你必须为正确的数据量选择正确的模型。 我提到过。 新网络的生存依赖于大量数据。 如果你的数据量很少，那么使用回归分析等更简单的模型会更有意义。 P 值黑客是我最讨厌的，我会在其他课程中多次谈到这一点，实际上是在你的数据分析原理中，因为这是我经常看到的，尤其是在论文阶段，可悲的是在很多学术论文中也是如此。 价值黑客意味着人们在寻找一些并不存在的东西。 他们拼命想找到一种关系，因为他们花了那么多时间和金钱在某些事情上，他们测试了这种关系，然后他们报告了一些微小的显著性，或者他们报告了一些类似于，嗯，这并不显著，但我们仍然认为它是。 这就是 不，这说不通。 意义是二元的。 要么有意义，要么没意义。 就是这样 它只是如实报告数字，然后你就跟着它走。 这也与上一点有关，即过度依赖理论和常识。 例如，你可以在经济学中看到这一点。 哦，我希望我的同事们都不看这个。 没有，没有。 他们真了不起。 我在几年前的一篇论文中就看到过这种情况，人们认为经济理论中的某些具体内容被认为是正确的，即使数据集与之不同，也一定是数据错了。 而不是模型。 显然，这种情况不仅发生在经济学领域，这只是我的一个研究领域，因为我读了这篇论文，它让我吃了一惊，但它也发生在史丹姆领域，实际上是在物理学领域，例如，你收集数据，你有这些关于宇宙可能如何运作的理论模型，然后你收集数据，数据告诉你可能不是，但他们认为这个模型看起来很美，而且还有诺贝尔奖什么的。 我也不知道。 所以数据一定是错的。 数据收集过程肯定有问题。 很多情况下确实如此。 在很多情况下，数据质量会影响结果，影响实际报告内容。 但在其他情况下，可能是经验证据指明了方向。 所以，让我们相信经验证据，而不是我们认为和相信的绝对真理。 好吧，我意识到我说了很多。 因为这是你的第一次演讲，我想让你再跟我多说几分钟。 五到十分钟，告诉我你对我们目前讨论的内容有什么看法。 这是否符合你们的期望？ 还有，对我所讲的内容有什么问题吗？ 有问题就问吧。 这就是我来这里的真正原因。 所以，我试图把材料。 你可以直接读这本书。 所以，我的工作就是把这些知识，变得易懂、可呈现、适用于你，并以一种适合你的方式呈现给你，这样我们就能一起从中学习。 我并不是在这里发明新知识。 我是在用一种希望有用的方式包装知识。 因此，得到你们的反馈和提出你们的问题对我来说非常重要。 如果你认为这是一个愚蠢的问题，它可能不是。 你的同事可能也有同样的疑问，但如果你不好意思在第三节课上提问，可以给我发电子邮件，或者在课间休息或课后问我，我同样很乐意与你一对一交流。 对课程结构有什么直接的问题吗？ 我告诉过你的任何事情？ 任何无关的问题？ 有的 我知道我们有考试。 是只有十二月的一次考试 还是也会有期中考试？ 只有一次考试 只有一次考试 所以，评估只是课程作业，我想很有可能在 11 月提交。 然后十二月初考试。 是的 一次作业 一次考试 没有期中考试 没有期中考试 我们会不会从每周一次变成每月一次？ 会很酷的。 是啊。 所以没有持续的评估。 所以在学期中没有持续的分级。 我们确实有电脑实验室，我想明天你会有更多的电脑实验室，我们会有一些练习单让你去做，你会从我这里得到解决方案和反馈。 但我不会给你们的作业打分，所以你们可以仔细检查我提供的解决方案和我在课堂上给出的反馈。 什么？ 什么？ 12 月初的 Rite Aid。 我真的想参加我的圣诞节。 周五你就知道了 是的，是的。 是的，是的。 然后，我想至少截止日期 为课件将是 将在本周五 只要它通过。 审核 考试时间 我不能告诉你 考试 我想我们有两周的时间。 因此，我建议您等待与调度任何你的假期，直到你有你的考试日期，因为他们可以深入到圣诞节。 我见过 12 月 23 日的考试。 所以，如果你认为你得救了，因为它是圣诞节前两天，你是不是安全的。 是的。 课程将如何 这门课程的内容配合课程？ 嗯。 是啊，这是个好问题。 所以，我想我们之前在讨论描述性和描述性与预测性的时候，你的一位同事已经提到了这个问题。 因此，数据分析的原因和原理是一门基础统计学课程，所以它将更偏向于描述性分析方向。 我们将涵盖假设检验、方差分析、Anova 等内容，以及描述性统计与论述的对比，后者更偏重于机器学习。 因此，我们不会像那门课程那样进行传统的统计，但显然它们会并行进行，所以你会看到一些重叠。 这门课程还与第二批学生一起上课。 所以你会遇到金融科技项目的同事。 别问我是哪个。 我想我最近改名字了。 金融科技项目之一。 所以你们的队伍非常庞大。 我想我们有115人左右。 这么多学生 只有五周的课程 是统计学速成班 是啊 Yeah. 还有别的事吗？ 退休了 这是你第一天上课 希望你还不累 因为还有很多课要上呢 你最喜欢的数据集是哪个? 这是你最喜欢的数据集？ 我最喜欢的数据集 这真是个棘手的问题。 目前我的研究兴趣之一是我之前提到的财务福利和财务健康。 最近，我正在使用加拿大金融消费者机构（FCA）的数据集，我们正在研究 Covid 19 期间人们的财务状况，这非常有趣，因为他们多年来一直在收集这类数据。 然后 Covid 发生了，所以他们改变了调查内容，但仍有一些变量保持不变。 因此，我们实际上可以看看加拿大人在大流行之前的财务状况发展，然后在大流行之后和现在的恢复阶段，他们又回到了那种定期更新的状态。 我认为这种时间上的发展对我来说非常有趣，尤其是因为其中还有空间上的因素。 因此，我可以看看空间上的不平等，以及加拿大不同地区比其他地区恢复得更快。 例如，你可以将其与社会经济弱势群体联系起来，这些群体在恢复过程中更加艰难。 因此，我认为这对我来说是一个非常有趣的数据集，因为它非常具有话题性，非常新颖，而且相对干净，这对我的工作总是很有帮助。 是的，你也可以问我关于我的研究的问题。 我很乐意回答。 不，你问完了。 我可以 我看到你开始收尾了 我知道了 你知道的 都不错。 所以，让我们总结了一下， 总结早了一点，今天 因为我们的休息时间有点短。 明天在你们的讲座和计算机实验室见 下周见。 如果你们还有其他问题，可以给我发电子邮件，我可以安排见面或通过电子邮件回答。 酷毙了 谢谢 再见 再见