分析 1988 年到 2008 年间的美国航班延误情况

陈子萌、黎思言、闫施、尹航

摘要

本文使用了 1998 年到 2008 年的每月起飞航班总数、每年起飞航班总数、每月延误航班总数、每年延误航班总数、每月延误航班延误平均时间等数据。基于这些数据，我们整体分析了航班的延误情况。然后我们统计了 1998 年到 2008 年的每年每个机场的起飞航班总数、起飞延误航班总数、机场所在城市和机场所在州等数据。基于这些数据，我们分州和城市分析了航班的延误情况。我们统计了美国各州拥有的机场数、总共起降的航班数、取消航班比例、出发延迟超过 10 分钟的航班比例，到达延迟超过 10 分钟的航班比例等指标，对美国 52 个州进行了聚类分析。最后我们统计了纽约肯尼迪机场 20 年来的每天航班起飞总数、起飞延误总数、降落飞机总数、降落延误航班总数和天气信息等字段。基于这些数据，我们使用随机森林分析天气数据对延误率的影响。

1 数据和方法

本文使用的了从 1988 年到 2008 年的飞机航班数据和纽约肯尼迪机场 1998 年到 2008 年的的天气数据。本文从数据库中下载了所有的飞机航班数据和所有的天气数据，然后将数据导入了本地数据库。在本地数据库中统计了 1998 年到 2008 年的每月起飞航班总数、每年起飞航班总数、每月延误航班总数、每年延误航班总数、每月延误航班延误平均时间等数据。基于这些数据，我们整体分析了航班的延误情况。在本地数据库中统计了 1998 年到 2008 年的每年每个机场的起飞航班总数、起飞延误航班总数、机场所在城市和机场所在州等数据。基于这些数据，我们分州和城市分析了航班的延误情况。在本地数据库中，我们统计了美国各州拥有的机场数、总共起降的航班数、取消航班比例、出发延迟超过 10 分钟的航班比例，到达延迟超过 10 分钟的航班比例等指标，对美国 52 个州进行了聚类分析。在本地数据库中统计了纽约肯尼迪机场 20 年来的每天航班起飞总数、起飞延误总数、降落飞机总数、降落延误航班总数和天气信息等字段。基于这些数据，我们使用随机森林分析天气数据对延误率的影响。

2 结果

2.1 整体航班延误情况

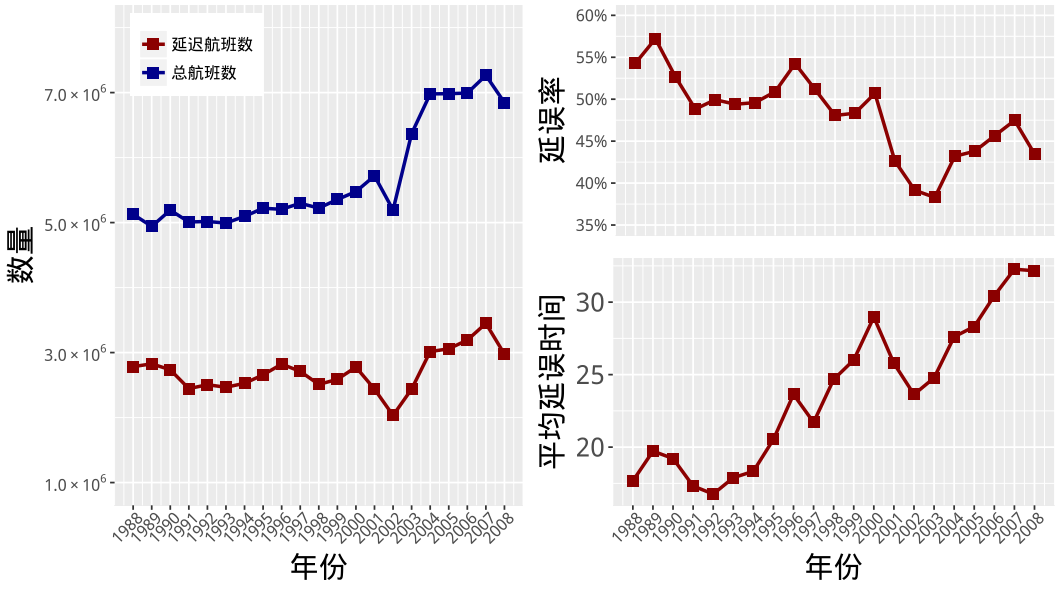

图 1: 分年度整体分析

20 年来，美国航班的延迟航班数、总航班数、延迟率和平均延迟时间如图 1 所示。延迟航班是指，航班的实际出发时间比计划的出发时间晚。延迟率是指，延迟航班总数和总航班数的比值。平均延迟时间是指，所有延迟航班的平均延迟时间，单位是分钟。从图中得出以下结论：

（1）20 年来，航班总数和延迟航班数的整体趋势是上升的。

（2）20 年来，延迟率整体趋势是下降的。上世纪 90 年代，每年有大于 50% 的航班出现延误。但是在 2000 年之后，延误率降低到 45% 以下，在 2003 年之后稍有波动。

（3）虽然延迟率有显著的下降，但平均延迟时间却上升了。上世纪 90 年代，平均延迟时间在 25 分钟以下，2000 年以后，平均延迟时间不断上升，目前已经超过了 30 分钟。

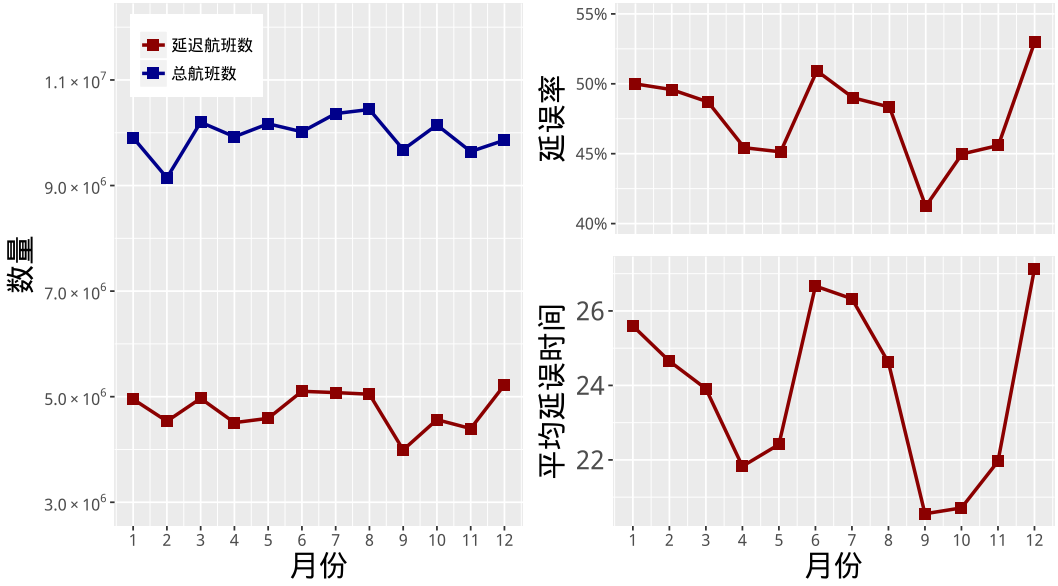

图 2: 分月份整体分析

每个月的延迟航班数、总航班数、延迟率和平均延迟时间如图 2 所示。从图中得出以下结论：

（1）12 月、1 月、6 月、7 月、8 月的航班总数和延迟航班数都比其他时间高。说明这段时间是乘坐飞机出行的高峰时期。

（2）在上述的高峰月份，延迟率都大于其他月份。

（3）在上述的高峰月份，延迟时间都大于其他月份。

2.2 各州航班延误情况

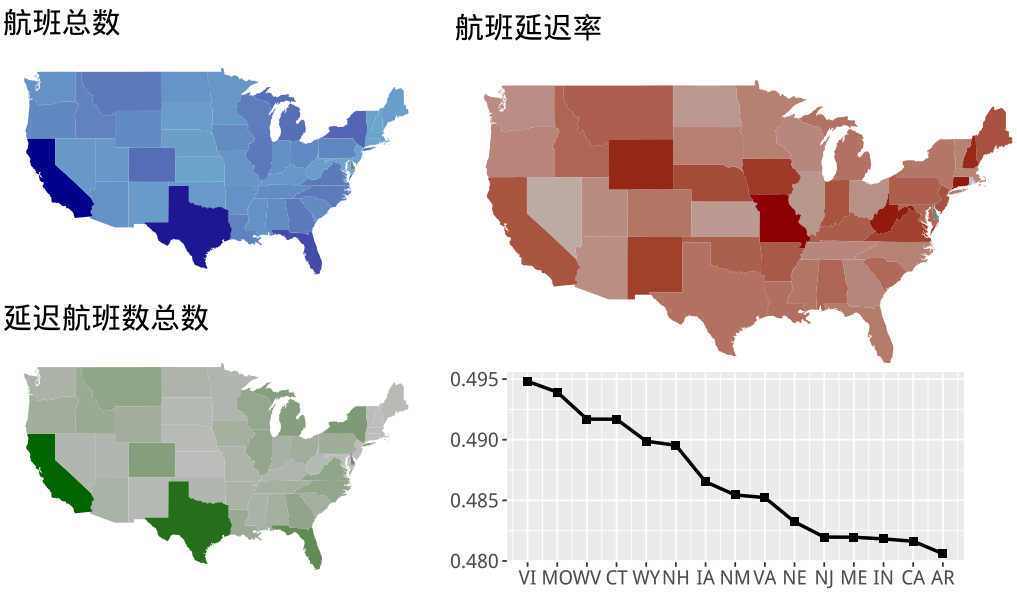

图 3: 各州航班延误情况分析

使用 2008 年的航班数据，统计了每个州的航班情况如图 3 所示。左上角颜色越深表示航班数量越多，左下角颜色越深表示延迟航班数量越多，右上颜色越深表示航班延迟率越高。右下标注出

了延迟率最高的 15 个州。从图中得到以下结论：

（1）航班总数和延迟航班总数最高的州是加利福尼亚州和得克萨斯州。加利福尼亚州是美国本土面积第二大州，人口第一大州和国内生产总值第一大州。德克萨市州是美国本土面积最大州，人口第二大州和国内生产总值第二大州。

（2）美国中部各州航班总数和延迟航班数不高，但航班延迟率相对较高。

（3）航班延迟率最高的 15 个州为：佛蒙特州、密苏里州、西弗吉尼亚州、康涅狄格州、怀俄明州、新罕布什尔州、爱荷华州、新墨西哥州、弗吉尼亚州、内布拉斯加州、新泽西州、缅因州、印第安纳州、加利福尼亚州、阿肯色州。这些州大部分位于美国中西部地区。

2.3 主要城市航班情况

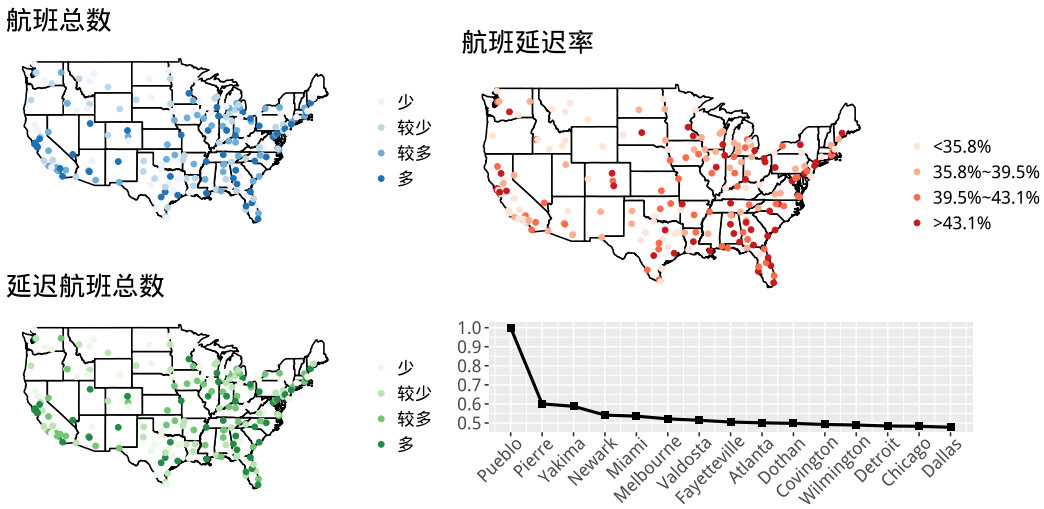

图 4: 主要城市航班情况

使用 2008 年的航班数据，统计了美国主要城市的航班情况如图 4 所示。左上角颜色越深表示航班数量越多，左下角颜色越深表示延迟航班数量越多，右上颜色越深表示航班延迟率越高。右下标注出了延迟率最高的 15 个州。从图中得到以下结论：

（1）美国东部城市、五大湖周围城市和西海岸加利福尼亚州城市群的航班数量和延迟航班数量较多。

（2）航班延迟率较高的城市集中在美国东部和中部。

2.4 使用机场信息聚类分析美国各州情况

我们统计了美国各州拥有的机场数、总共起降的航班数、取消航班比例、出发延迟超过 10 分钟的航班比例，到达延迟超过 10 分钟的航班比例等指标，对美国 52 个州进行了聚类分析。聚类方法使用的是 K 均值聚类法，迭代次数为 20 次，为了确定适合的分类个数 K，把 K 可能的取值范围设定到 2-7，分别计算 K 的各个取值下的轮廓系数以及组间方差和总方差之比值，对 K 在各取值的情况下 K 均值聚类的效果作出评价。

表 1: k 均值聚类分析

k 的取值	组间方差	轮廓系数
2	0.51776	0.7651
3	0.64168	0.2602
4	0.71223	0.2656
5	0.76510	0.2694
6	0.79325	0.2626
7	0.80487	0.2606

由表 1 可知，当 K 的取值从 2 变为 3 时，轮廓系数剧减，因此，首先判断 K 的最优取值为 2。但是，当 K=2 时，根据聚类结果可得，两类中机场数分别为 2 个和 51 个，也就是说，绝大多数机场都被分为一类了，并不能够达到分类的目的，因此，再次观察表格，发现当 K=5 时轮廓系数达到了最大，而组间方差/总方差在这之后随 K 取值增加的提升幅度也很小了，因此，把 K=5 作为比较理想的 K 取值。当 K 取 5 时，组间方差与总方差的比值为 76.6%，所有的地区被分为 5 个类，每一类中的地区数分别为 18，8，12，13，2.

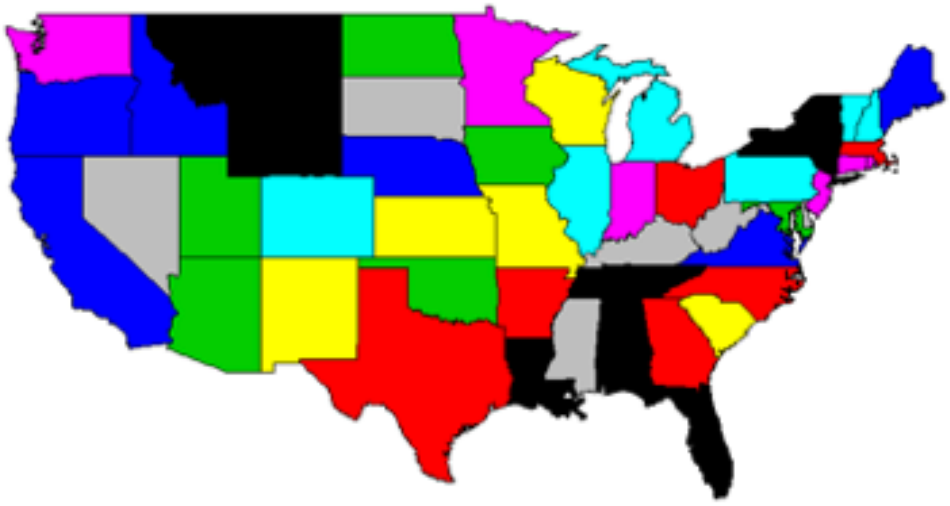

图 5: 聚类分析结果

聚类的结果如图 5 所示。类别 1 的地区特点为拥有最多的机场数量和总起降航班数，取消航班的比例也是最低的；类别 2 的地区，拥有的机场数量较少，取消航班的比例与降落延迟航班的比例也很低；类别 3 的地区拥有的机场数量较少，取消航班的比例较低，还拥有最高的起飞延迟比例和最低的降落延迟比例，类别 4 的地区则拥有最少的机场数，最低的总起降航班数，最高的取消航班比例和降落延迟航班比例，以及最低的起飞延迟比例；类别 5 的地区拥有的机场数量较多，取消航班的比例也较高。

2.5 使用天气信息预测肯尼迪机场的航班延误率

由上文可知，美国航线的延误情况不容乐观。动辄 40% 50% 的延误率不仅影响了人们的出行，更对美国的航空运输行业产生了负面的影响。航空延误的现象出现的原因多种多样，天气原因是造成航空延误的主要原因之一。出发地机场天气状况不宜起飞；目的地机场天气状况不宜降落；飞行航路上气象状况不宜飞越等等都可能导致航空延误。本文主要研究出发地天气状况不宜起飞或者目的地天气状况不宜降落导致的延误。

本文选取的研究对象是纽约约翰 · 菲茨杰拉德 · 肯尼迪国际机场。该机场是纽约市的主要国际机场，是全世界最大机场之一。机场于 1942 年始建，1948 年 7 月 1 日首次有商业航班，并于 7 月 31 日正式命名为“纽约国际机场”。1963 年 11 月 22 日美国总统约翰 · 肯尼迪遇刺身亡，12 月 24 日机场改名为“约翰 · 菲茨杰拉德 · 肯尼迪国际机场”以纪念这位先总统；随后，机场的国际航空运输协会机场代码更新为 JFK。肯尼迪国际机场设有 9 个客运航站楼，各航站以 U 形格局围绕机场中心区域的停车场、酒店、供电设施等设施。AirTrain 和道路均接驳各个航站。JFK 占地 4,930 英亩，其中包括中央航站区的 880 英亩。约 37,000 人在 JFK 工作。机场为纽约 - 新泽西州大都市区贡献约 373 亿美元的产值，同时创造了约 256,000 个工作机会，提供了 134 亿美元的薪资。JFK 还是世界领先的国际航空货运中心之一。机场提供面积近 400 万平方英尺的的货物仓库和办公空间。整个航空货运区都被指定为外贸区。JFK 通过长距离、直达和不间断的空运，为世界主要的航空货运市场服务。在本文中，我们使用肯尼迪机场的数据分析天气对肯尼迪机场的航班延误率的影响。

2.5.1 描述统计

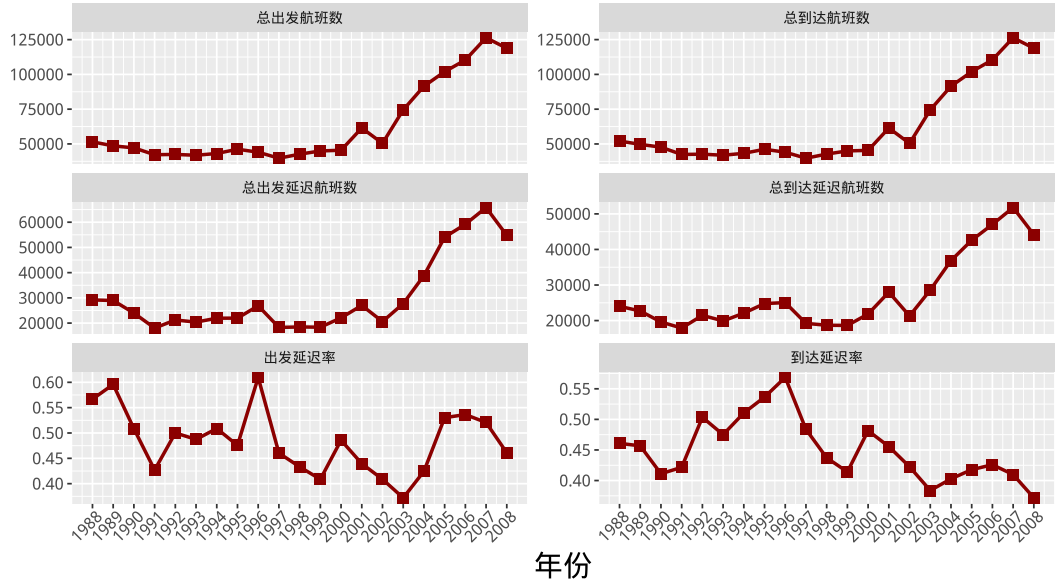

图 6: 肯尼迪机场基本情况

纽约肯尼迪机场 20 年来的起飞航班和降落航班的基本情况如图 6 从图中可以得出以下结论：

（1）2000 年以来，肯尼迪机场起飞的航班和到达的国内航班数量都显著增加。从一年 5 万趟国内航班的水平上升到一年 12.5 万趟国内航班。增长了约 1.5 倍。

（2）2000 年以来，肯尼迪机场出发的延迟航班和到达的延迟航班数量都显著上升。从一年 2 万班延迟航班的水平上升到一年约 5.5 万班。

（3）肯尼迪机场的出发航班和到达航班的延迟率都呈现出先上升后下降的变化趋势。

2.5.2 使用天气信息预测起飞航班的延误率

本文使用 1988 年到 2008 年每天纽约肯尼迪机场的天气信息预测从肯尼迪机场起飞的航班的延误率。

本文选取的自变量特征有。"Year": 年; "Month": 月; "DayofMonth": 日; "Maxtemp": 最高温度; "Meantemp": 平均温度; "Mintemp": 最低温度; "Maxdewpoint": 最高露点温度; "Meandewpoint": 平均露点温度; "Mindewpoint": 最低露点温度; "Maxhumidity": 最高湿度; "Meanhumidity": 平均湿度; "Minhumidity": 最低湿度; "Maxsealevelpre": 平均海拔高度: "Meansealevelpre"; 最低海拔高度: "Minsealevelpre"; 最高能见度: "Maxvisibility"; 平均能见度: "Meanvisibility"; 最低能见度: "Minvisibility"; 最高风速: "Maxwindspeed"; 最低风速: "Meanwindspeed"; 降水量: "Rainfall"; 云量: "Cloudcover"; 天气状况: "Events"; 风向度数: "Winddirdegrees" 等 24 个自变量。因变量是起飞航班的延误率，本文将起飞航班延误率小于 50% 的天分为一组，将起飞航班延误率大于 50% 的天分为另一组。

自变量中的缺失数据的处理办法如下，最高温度、最低温度和平均温度中的缺失数据删除，一共删除了 460 条缺失数据。最高能见度、最低能见度和平均能见度中有 12 个缺失数据，采用中位数插补。风向度数中有 7 个缺失数据，采用中位数插补。天气状况中的空缺数据，采用"orther"水平补齐。处理完毕后，一共有 7208 个有效数据。在延误率较低的组中有 4026 个数据，在延误率较高的组中有 3182 个数据。

本文将上述数据按照 75% 的训练集和 25% 的测试集进行划分，对训练集样本建立了 300 棵 CART 分类树的随机森林模型。输出预测结果和自变量的重要程度。重要程度如图 7 。预测准确率如表 2。可见预测准确率为 75.42%. 在延误率较低的组中，预测准确率是 79.1%. 在延误率较高的组中，预测准确率是 70.75%. 对飞机起飞延误影响最大的天气因素是天气状况，其次是最低能见度。

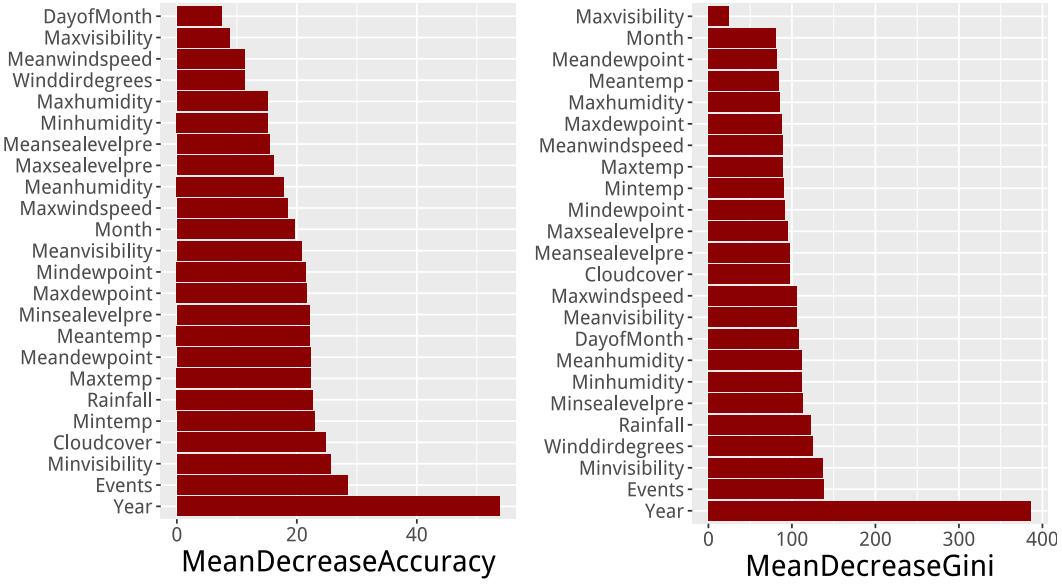

图 7: 使用天气信息预测起飞航班的延误率的变量重要性

表 2: 使用天气信息预测起飞航班的延误率的预测混淆矩阵

	(0,0.5]	(0.5,1]
(0,0.5]	788	207
(0.5,1]	236	571

2.5.3 使用天气信息预测降落航班的延误率

降落航班的自变量和因变量选取过程和缺失值处理过程和起飞航班的处理过程类似。在延误率较低的租中有 4888 个数据，在延误率较高的组中有 2320 个数据。本文依旧划分 75% 的训练集和 25% 的测试集，使用 300 棵分类树的随机森林模型估计降落航班的延误率。输出预测结果和自变量的重要程度。重要程度如图 8 。预测准确率如表 3。可见预测准确率为 78.02%.

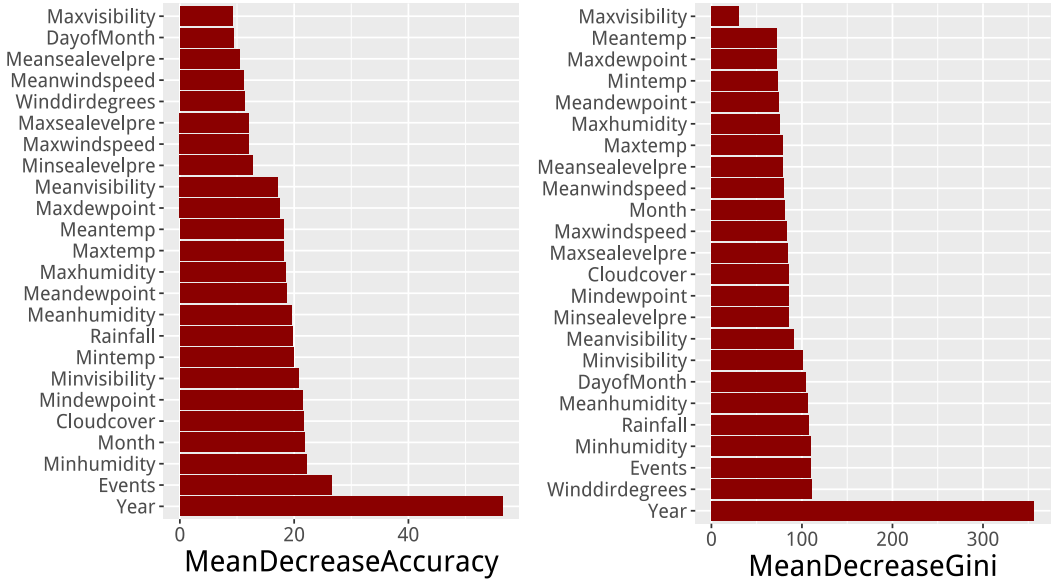

图 8: 使用天气信息预测降落航班的延误率的变量重要性

表 3: 使用天气信息预测降落航班的延误率的预测混淆矩阵

	(0,0.5]	(0.5,1]
(0,0.5]	1078	123
(0.5,1]	273	328

从图 8 中可知，对飞机降落延误影响最大的天气因素有：天气状况、最低湿度、云量等。从预测效果看来，对延误率较小的组的预测效果较好，但是对延误率较大组的预测效果不好。

这是因为，在降落航班中，延误率较高的数据和延误率较低的数据个数并不平衡。对于不平衡数据一般采用对少数组过抽样或对多数组欠抽样的方法提高对少数组的预测精度。本文采用过抽样的方法解决不平衡数据的分类问题，本文采取的不平衡数据处理方法是对少数组过抽样的方法。我们在训练集中对少数组过抽样了 1000 个样本。过抽样之后的预测结果和变量重要程度如图 9 和表 4。

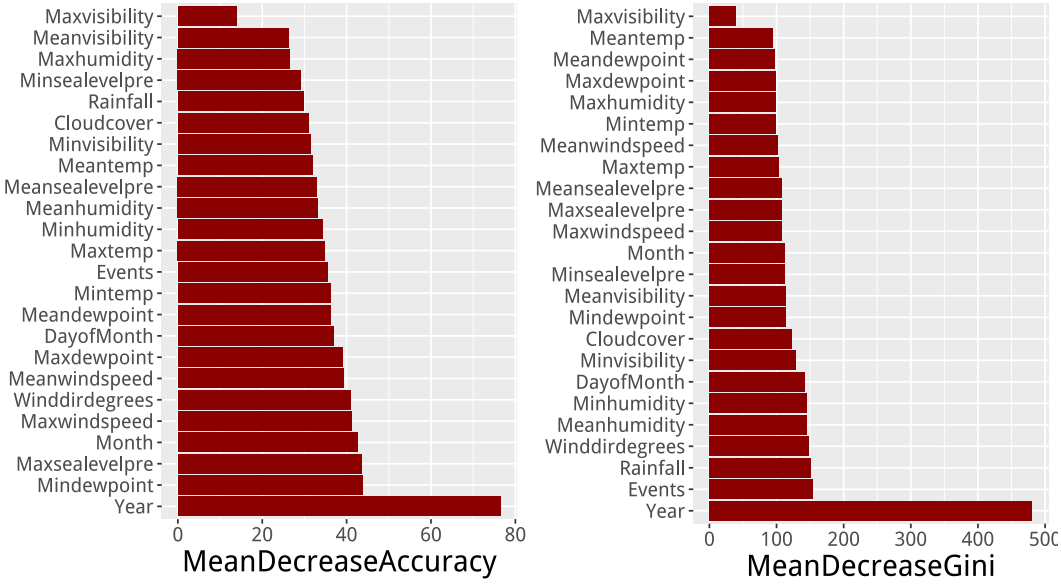

图 9: 使用天气信息预测降落航班的延误率的变量重要性 2

表 4: 使用天气信息预测降落航班的延误率的预测混淆矩阵 2

	(0,0.5]	(0.5,1]
(0,0.5]	1035	166
(0.5,1]	234	367

从图 9 中可知，对飞机降落延误影响最大的天气因素有最低露点温度、最高海平面高度、最高风速等。预测准确率为 77.80%. 总体预测准确率下降了，但是对于少数组的预测准确率上升了。准确率从 57.58% 上升到 61.06%.