再见，航班延误

——基于美国交通部运输统计局航班数据分析

strength library and muscle laboratory（第七组）

陈立国 樊可 韩思越 黄永晟 杨远琨 俞铖昊

一、背景

随着人们生活水平的提高，生活节奏的加快，越来越多的人倾向选择更为快捷的航班出行方式。然而，出于天气、空管、航空公司等种种原因，航班大概率延误的问题，始终为航班乘客投诉和关注的中心话题。美国交通统计局数据显示，美国航空公司 2018 年延误的航班超过 100 万架次，相比 2017 年增加 25%。并且，平均每五架次航班，即有一架次的航班延误。

对差旅党而言，一次延误可能意味着一场重要演讲的缺席，一次重要谈判的取消。对旅行团而言，一次延误可能会促成旅行项目的取消，游客的负面抱怨与投诉，进而影响旅行团口碑与业绩。加州伯克利大学调查显示，美国国内每年，航班延误造成的总成本高达 329 亿美元，其中有一半的成本，不得不由乘客以各种间接方式承担。而《航空公司航班延误的经济成本》计算显示，如果将航班延误减少 30%，美国社会的净福利将增加 385 亿美元。可见，航班的延误对个人对社会损失巨大，如何尽可能规避航班延误，或减小航班延误的可能性，已是迫在眉睫的问题。

我们对美国交通部运输统计局数据 2018 年的美国随机抽取 100 架飞机的航班飞行数据进行分析，以航班是否延误为因变量，其他各类因素为自变量建立模型进行相应分析。差旅党能够根据我们的分析选择延误率更低的航班，并根据航班是否可能延误来调整个人时间规划安排，尽可能规避因航班延误导致的重要会议谈判迟到等个人损失。旅行团可以选择延误率更低的航班提升游客对旅途满意程度，并提前预估所定航班可能的延误情况安排行程，将航班延误对旅游项目游客体验的影响降至最低。

二、数据说明

本案例随机挑选 2018 年 100 架飞机航班数据，删除因航班取消，航线改道变量缺失数据，筛除实际情况下无法预知考虑的“实际起飞时间”“起飞延误”等自变量，最终保留 125311 条数据。

后续建模中，我们将“到达延误”大于 14 分钟数据条标为延误数据，其他数据条标为未延误数据作为因变量。根据飞机机型信息爬取飞机制造商作为自变量，定义始末机场距离与系统预定飞行时长比值（CRS_ELAPSED_TIME/Distance）为计划空速作为自变量，将“飞机制造商”自变量按规模分类。预处理后用于下文描述分析与模型分析的变量情况如表 1 所示。

表 1 变量表

变量类型		变量名	解释	详细说明	取值范围与备注
因变量		delay	是否延误	定性变量 共 2 个水平	TRUE, FALSE
自变量	航线信息	Carrier_Size	航空公司大小	定性变量 共 3 个水平	Large, Medium, Small
		COMPANY	飞机制造商	定性变量 共 5 个水平	A(空中客车),B(波音),C(庞巴迪),E(巴西航空工业),M(麦克唐纳·道格拉斯)
		origin_airport	出发机场	定性变量 共 3 个水平	Busy, Medium-size, Small
		dest_airport	到达机场	定性变量 共 3 个水平	Busy, Medium-size, Small
		Distance	航班航程	定性变量 共 4 个水平	<250, 250~550, 550~1200, >1200 单位：英里
		flight_speed	计划空速	连续变量	0.95 ~ 13.57 单位：英里/分钟
	时间信息	Season	季度	定性变量 共 4 个水平	Spring, Summer, Autumn, Winter
		Week	星期	定性变量 共 7 个水平	周一~周日
		CRS_AT_Hour	计划到达时间	定性变量 共 3 个水平	AH0-4, AH5-16, AH17-23
		CRS_DT_Hour	计划出发时间	定性变量 共 3 个水平	DH0-4, DH5-16, DH17-23

三、描述分析

本案例关注的因变量是航班是否延误。由图 1 航班延误柱状图中可以看出，此次分析的数据中延误的航班总数超过两万次，几乎接近航班总数的五分之一。可见航班延误概率比例之大，果然是差旅党，旅行团相当头疼的问题。

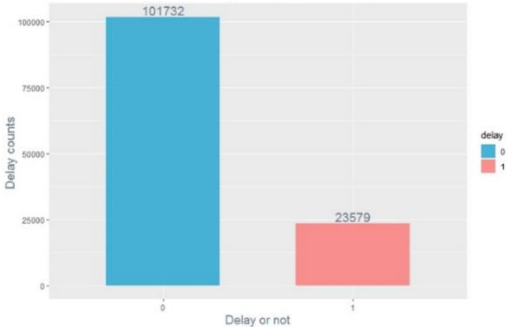

图 1 未延误航班与延误航班量柱状图

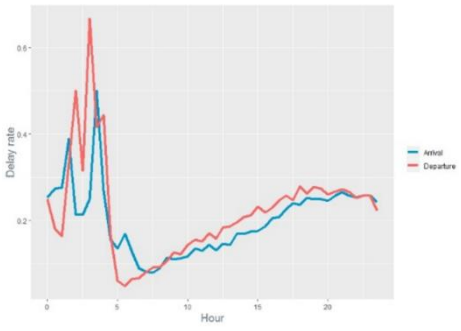

图 2 航班起降时间与延误率折线图

我们接着来看各个自变量对于航班延误的影响。先从时间信息中的航班起降时间入手。如图 2 所示，凌晨时段起降的航班延误率高且波动大，推测波动源于该段航班数据量少导致估计不稳定，延误率高源于该段疏于管理易造成航班延误。从清晨到深夜，航班的延误率逐渐增高，猜测原因是清晨到深夜航班数量多，机场起降道安排较满，一旦某航班延误，大概率会占用该起降道下一架航班，由此越晚起降的航班，起降道受到其他延误航班占用的概率越大，延误率也越大。

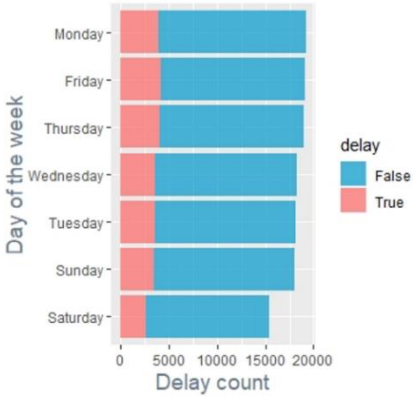

图 3 各星期航班量柱状图

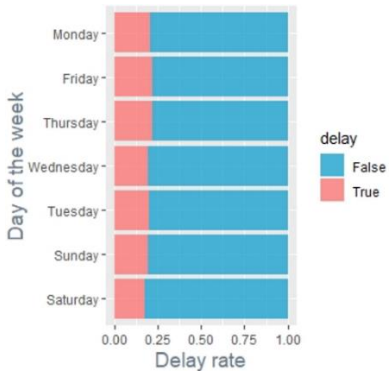

图 4 各星期延误率柱状图

然后考察星期与航班延误率的关联。比较图 3 和图 4 可以发现，航班总量最少的周六，延误率最低，而航班总量偏多的周一、周四、周五，延误率均明显偏高。猜测多数航空差旅项目与旅行项目安排周一启程，周四、周五返程导致航班总量偏多，也导致这几天航道占用更满，航班延误更易占用后续航班的起降道，致使后续航班的延误。

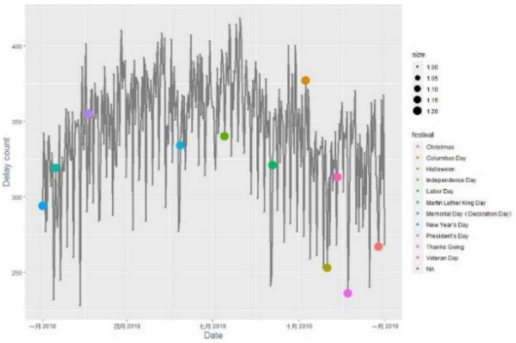

图 5 日期与延误次数折线图

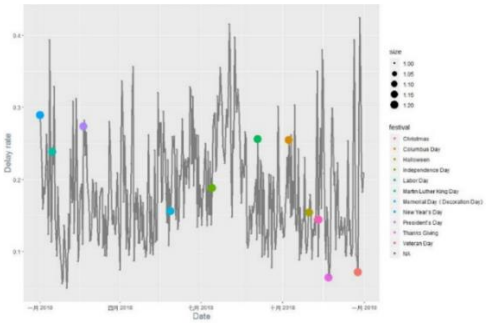

图 6 日期与延误率折线图

接着考察日期与航班延误率的关联。图 5 与图 6 分别反映了 2018 年各天的航班延误数量与延误率（标记点为节日信息）。总体上航班在春夏季延误数量偏多，秋冬季偏少。在感恩节与圣诞节前后几天，延误数量与延误率均较高，而在这些节日当天，航班延误数量与延误率均骤降到极低点。猜测重大节日前后几天，出游旅客增加导致航班增加航道占用更满，航班延误对后续航班的按时起降影响更大，因此延误率升高。而节日当天，各机场为防止航班延误造成不良社会影响而加强航班管理与调控，使得当天延误率骤降。而部分其他节日，如元旦总统日和劳动日，延误率均骤升到较高点，推测这些小型节日并未受到机场重视，也即没有将强管理。通过上述两表，我们可以为我们在节假日飞机出行作出提前规划，尽量避开航班延误的高峰期。

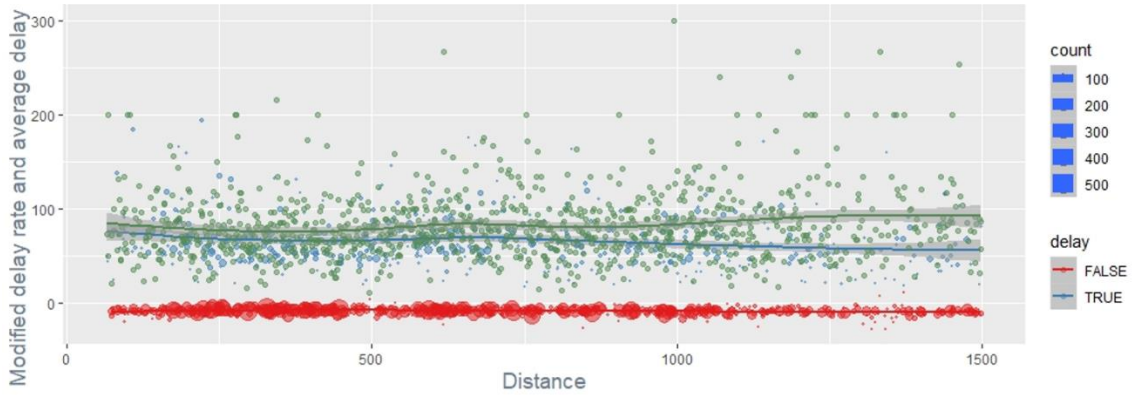

图 7 距离与延误时长散点图

分析完时间信息，我们再来分析航线信息对航班延误的影响。首先是始末机场的距离。如图 7，蓝色与红色点分别代表不同始末机场距离下，平均延误时长与平均提前到达时长。随着机场距离的增大，平均延误时长减少，平均提前到达时长增大。绿点代表了不同始末机场距离下的航班（400 倍）延误率，航线的长度对航班延误率影响并不明显。

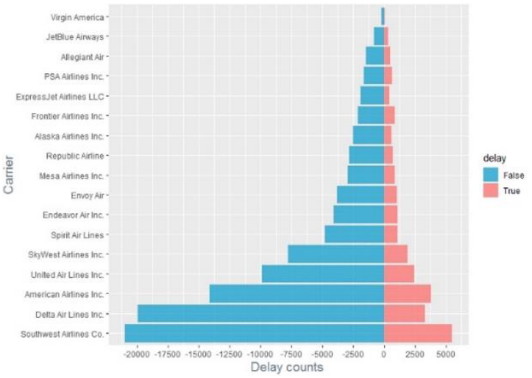

图 9 各航空公司未延误与延误量柱状图

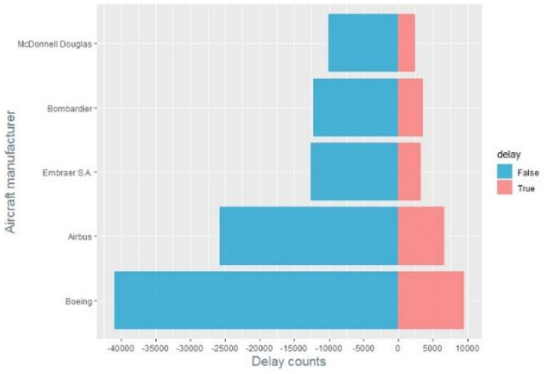

图 10 生产制造商未延误与延误量柱状图

最后，我们再来考虑航空公司与生产制造商对延误的影响。如图 9 所示，大型航空公司中，美国航空的航班延误率特别高；而在小型航空公司中，边疆航空公司航班延误率也明显较高。而图 10 说明，作为民航制造两大巨头空客和波音，空客的延误率远高于波音。在这 5 家制造商中，庞巴迪公司的飞机航班延误率最高。由此我们建议无论是差旅党还是旅行团，选择航班的过程中尽可能选择波音航空公司，而规避空客和庞巴迪两家民航制造商。

四、逻辑回归模型

首先，我们使用逻辑回归模型进行分析。图 11 展现了部分变量的系数比较，受篇幅有限，不再展现各变量系数，结果如下：

1. 航空公司：航空公司规模越大，延误率越低。其中中型航空公司比大型航空公司高了 11%，小型航空公司比大型航空公司高了 45%。因此建议差旅党有条件的情况下更多地选择大型航空公司航班，能够更多地减少航班延误带来的时间损失与计划安排变动。
2. 季节：夏季延误率较高，延误率发生比较春季高了 45%。由此建议旅行团在夏季安排行程过程中为航班延误预留更为充足与宽裕的时间，尽量避免因为航班延误而导致的旅行计划搁浅与不良游客体验。
3. 星期：星期六延误率特别低，延误率发生比航班较多的工作日（周一、周四、周五）减少约 19%。因此建议差旅党与旅行团尽可能选择周六的航班。
4. 计划出发与到达时间：对比上午 5 点到下午 16 点的延误率与下午 16 点到晚上 23 点的延误率，计划出发时间前者比后者低了 44%，而计划到达时间前者比后者低了 35%，因此强

烈建议差旅党选择机票时尽可能选择早晨起飞与到达的机票，由此不仅仅能够减少航班延误的可能性，还能够错开来回机场的上班高峰时间。

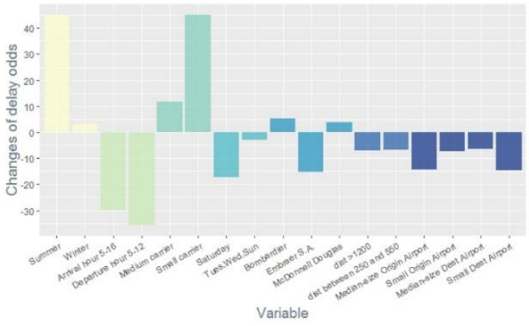

图 11 逻辑回归自变量回归系数柱状图

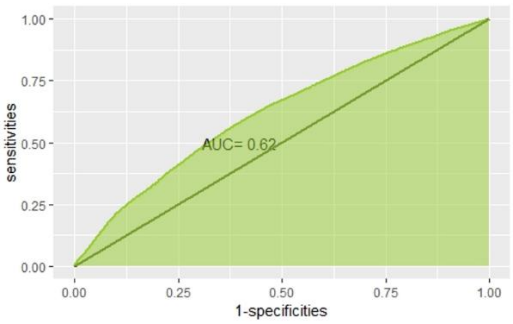

图 12 逻辑回归 ROC 曲线图

由于经过 AIC，BIC 变量选择，十折交叉验证后，图 12 中 ROC 曲线与坐标轴围成的面积（AUC）为 0.62。即考虑对一架延误的航班与一架未延误的航班，预测其中哪一架航班会延误。如果直接使用逻辑回归模型，预测成功的概率为 62%，略高于随机猜测正确概率 50%，因此能够为航班延误预测提供一定程度的参考。

五、回归树模型

我们分别使用的传统的 CART 决策树模型，随机森林模型以及提升树模型对该数据集进行分析。使用十折交叉验证，对训练集使用欠采样的方式进一步采样得到“延误”与“不延误”比例相对平衡数据，使得训练集不再倾向数据多的“不延误”类别，再分别使用 CART 决策树，随机森林，提升树模型进行分析。

如表 2 所示，尽管“全部不延误”的预测方式预测准确率最高，但考虑到实际问题中差旅党和旅行团更为关注的是能否确保模型给出供选择的航班是否会发生延误，即确保模型能够给出一定数量的非延误航班（阳性率）的情况下，模型预测非延误航班中实际非延误航班占比（查准率）尽可能高。由此比较下，决策树模型最能够在确保阳性率不太小的情况下，具备最高的查准率。能够具备更高的效果。

因此，我们最后选择使用该决策树模型，对航班是否延误进行预测。受篇幅限制，不再展示大批模型预测数据的最终预测结果。举例说明，某差旅党希望预测 2019 年 11 月 26 日早晨 7：45 分起飞，西南航空公司，空客生产商的航班是否延误，只需将该航班数据输入模型当

中，即输出该航班“不会延误”的结果，即可作为差旅商的一条选择。

表 2 CART 决策树，随机森林，提升树准确率，查准率，阳性率表

	准确率	查准率	阳性率
全都不延误	0.875	0.875	1.000
决策树	0.438	0.915	0.377
随机森林	0.445	0.902	0.397
提升树	0.412	0.897	0.365

之后，我们再借助随机森林模型，查看各变量的影响程度。如图 13 所示，计划空速对航班是否延误的影响最为明显。推测计划空速越小，航班飞行过程中所允许的调整时间余量越大，由此越能够在航班飞行过程中利用调整时间余量，解决天气，航空系统等问题造成的时间耽搁，因此对整个航班延误是否延误的影响作用最大。

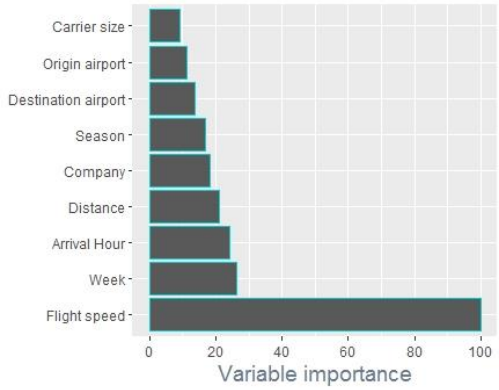

图 13 随机森林变量重要性柱状图

六、总结

为分析不同因素对航班是否延误的影响，为差旅党，旅行团等航班常客提供航班选择参考，我们对美国交通部运输统计局数据 2018 年航班数据使用逻辑回归模型，CART 决策树模型，随机森林模型以及提升树模型分别进行分析。

从分析结果中可以看出，航线信息方面，大规模航空公司往往比小规模航空公司延误率更低，波音的延误率相对低于空客和庞巴迪等制造商。时间信息方面，一天内从清晨到深夜延误率往往不断增大，一周内周六延误率最低，周一，周四，周五延误率最高，一年内夏季往往比

其他季节延误率更高，节日前后往往也会有更高的延误率。

由此，我们向差旅党和旅行团分别提出如下建议：

对于差旅党：

1、选择航班航线时，尽可能选择计划空速较低的航班，选择大规模航空公司航班与波音公司航班，由此能够有效地降低航班延误的可能性，有效减少因为航班延误带来的等待时间浪费，原定计划安排调动等不必要的损失。

2、选择航班日期时，建议尽量避开航班高峰节日时间，避开航班高峰的周一、周四、周五三天而尽量选择周六的航班。

3、选择航班时间时，建议选择清晨五点到八点的航班，由此不仅仅能够有效减少航班延误的可能性，还能够错开来回机场的上班高峰时间。

对于旅行团：

1、选择航班航线时，尽可能选择计划空速较低的航班，尽量避免选择小型航空公司，避免空客和庞巴迪两家民航制造商的机型，通过这样的方式来减小航班延误的可能性，从而降低由于航班延误带来的旅客不满与游程变动。

2、时间安排方面，尽量选择避开航班高峰的节假日和工作日时间的航班，鼓励组织清晨白天的航班出游，由此更易于减少航班延误概率，控制旅行安排时间。

附：小组分工情况

变量预处理：全员讨论变量处理方式，韩思越负责代码

描述分析：全员讨论呈现形式，黄永晟负责代码

模型分析：樊可负责逻辑回归模型，俞铖昊负责回归树模型

分析报告：全员讨论整合细节，杨远琨负责最终整合

小组汇报：陈立国负责 PPT 制作，杨远琨与陈立国负责汇报