数据挖掘
2019 年秋
课程大作业

基于机器学习分类模型的航班延误分析与预测

何福铿

中山大学数学学院应用统计学

目 录

表目录

摘 要

准确在飞机起飞前预测航班是否会延误可以让出行旅客更好地规划出行方式，以及给航空公司或者保险公司提供重要依据以提出航空延误险。为了在飞机起飞之前预测航班是否延误三小时以上，我们不仅考虑了航班的基本信息，还以机场的特情，天气等重要信息作为预测的依据。经过特征工程，我们挖掘出了飞机编号，航空公司，起飞时间段，机场，航班编号，计划飞行时间，天气，最高气温，最低气温，特情内容数量，特情紧急程度，月份等重要的特征。利用这些特征，我们以五个非线性模型作为备择模型，在训练集上进行模型训练，在测试集上进行验证，得到单模型的最高 auc score 为 0.6879。我们对这些模型进行模型融合，得到融合后的模型最高的 auc score 为 0.6902。

关键字：特征工程，ROC 曲线，非线性模型，模型融合

1 背景描述

随着国内民航的不断发展，航空出行已经成为人们比较普遍的出行方式，但是航班延误却成为旅客们比较头疼的问题。台风，雾霾或飞机故障等因素都有可能导致大面积航班延误的情况。飞机延误给旅客出行带来很多不便，所以我们要让利用大数据在航空领域发挥作用，在计划起飞前 2 小时预测航班是否会延误 3 小时以上，并给出延误 3 小时以上的概率，让出行旅客更好地规划出行方式。

另外，保险公司或者航空公司可以根据飞机延误 3 小时以上的概率，推出航空延误险，并且设置合理的赔偿额度（航班延误险，是指投保人根据航班延误保险合同规定，向保险人（保险公司）支付保险费，当合同约定的航班延误情况发生时，保险人依约给付保险金的商业保险行为）。

2 项目计划

本项目旨在准确在飞机起飞前预测航班是否会延误，为此收集了 2015 年 5 月到 2017 年 5 月历史航班动态起降数据，天气，以及特情数据，以提供一个可靠的预测模型。并将该模型用于实际的航班延误预测，检验该模型的可靠性以及提供进一步改进模型的依据。

3 分析数据集

3.1 数据描述

数据包含 2015 年 5 月到 2017 年 5 月历史航班动态起降数据，历史城市天气表，机场城市对应表以及历史机场特情表。具体变量说明如 表 1 表 2 表 3 表 4所示：

表 1. 城市天气

变量名称	变量分类	变量描述	变量说明	变量备注
城市	预测自变量	中文文字名词	多分类变量	如“上海”“广州”
天气	预测自变量	中文文字短语	文本变量	如“多云”“阴转小雨”
最高 / 最低气温	预测自变量	连续变量	数值变量	-29～47 摄氏度
日期	预测自变量	字符串	日期变量	如“2015/5/1”

表 2. 航班动态数据

变量名称	变量分类	变量描述	变量说明	变量备注
出发 / 到达机场	预测自变量	由三个大写英文字母组成	多分类变量	如“HGH”“DLG”
航班编号	预测自变量	数字和大写字母混合	多分类变量	如“CZ6328 “
计划起飞时间/计划到达时间/				
实际起飞时间/实际到达时间	可用于计算预测目标	时间戳	数值变量	如“1.45E+09”
飞机编号	预测自变量	飞机的编号	多分类变量	如 1,2,3
航班是否取消			二分类变量	如正常，取消

表 3. 特情

变量名称	变量分类	变量描述	变量说明	变量备注
特情机场	预测自变量	由三个英文字母组成	多分类变量	如“HGH”“DLG”
收集时间/开始时间/结束时间	预测自变量	时间戳	数值变量	如“1.45E+09”
特情内容	预测自变量	中文文字描述	文本变量	如“已出现大面积延误”

表 4. 机场城市对应表

变量名称	变量分类	变量描述	变量说明	变量备注
机场编码	预测自变量	由三个大写英文字母组成	多分类变量	如“HGH”“DLG”
城市名称	预测自变量	中文文字名词	多分类变量	如“上海”“广州”

3.2 数据清洗

（1）城市天气数据集中，清除掉机场城市对应表中没有的城市，并且清除掉错误的数据行。如图 1所示。

312375	罗萨里奥	多云	14	28	2016/3/23
312376	伊丽莎白洋	多云转小雨	18	21	2016/3/23
312377	哥伦布	小雨	17 11	17 11	2016/3/23
312378	山打根	小雨	26	26	2016/3/23
312379	彭萨科拉	阴转小雨	22 20	22 20	2016/3/23
312380	印第安纳波	阴转小雨	17 11	17 11	2016/3/23
312381	亚罗士打	小雨	25	25	2016/3/23

图 1. 数据乱序（1）

（2）特情数据集中，清除掉机场城市对应表中没有的机场代码，机场代码大小写统一，并且清除掉错误的数据行。如图 3所示。

5045	jgd	2016-04-(2016-04-(2016-04-(	机场开放	
5046	can	2016-04-(2016-04-(2016-04-(	受清明小长假影响，目前（4月2号）通往广州机场高速部分路段出现拥	
5047				
5048		2、尽量搭乘地铁前往机场，地铁三号线北延线可直通航站楼；		
5049				
5050		3、需自驾前往机场的旅客，如遇塞车请避开机场高速，绕行106国道来机场；此外，也可经广花公路(S114)		
5051	yie	2016-04-(2016-04-(2016-04-(	机场开放	
5052	nbs	2016-04-(2016-04-(2016-04-(	跑道积冰	
5053	axf	2016-04-(2016-04-(2016-04-(	道面积冰	
5054	inc	2016-04-(2016-04-(2016-04-(	目前，银川机场有雾，能见度低，期间进出港航班可能会受到影响。	

图 2. 数据乱序（2）

（3）机场城市对应表中，删除掉缺失的数据行。如图 3所示。

150	JNZ	锦州
151	LXI	
152	SHF	石河子

图 3. 数据缺失

3.3 数据集合并和初步处理

我们根据以下规则合并和处理上述四个数据集以便于进一步的特征分析：

- 根据机场城市对应表以及城市天气表，可以查出出发机场的天气，气温等数据。
- 根据特情的起始时间和终止时间，航班动态数据中的起飞时间，可以查出在飞机起飞的时刻，特情的数量以及内容。
- 根据计划起飞时间和实际起飞时间计算出延误时间以及计算出预测目标变量**是否延误三小时以上**（为表述方便，下文的航班延误为延误三个小时以上）。

合并后的数据集如表 5所示

表 5. 城市天气

变量名称	变量分类	变量描述	变量说明	变量备注
出发 / 到达机场	航班宏观背景	由三个大写英文字母组成	多分类变量	如“HGH”“DLG”
出发城市		中文文字名词	多分类变量	如“上海”“广州”
航班发生日期		时间戳	数值变量	如“1.45E+09”
航班发生天气		中文文字短语	文本变量	如“多云”“阴转小雨”
最高 / 最低气温		连续变量	数值变量	-29～47 摄氏度
特情内容	突发情况	中文文字描述	文本变量	——
计划起飞 / 到达时间	飞机自身情况	同“航班发生日期”	数值变量	如“1.45E+09”
实际起飞 / 到达时间		同“航班发生日期”	数值变量	如“1.45E+09”
飞机编号		飞机的编号	多分类变量	如 1,2,3
航班编号		数字和大写字母混合	多分类变量	如“CZ6328 “
是否延误三小时以上	结果	“TRUE”延误三小时以上,“FALSE”正常起飞	二分类变量	无延误或者延误三小时以下约占 95%

其中，我们把航班取消的数据以及出现缺失，乱序的数据删除了。删除的原因如下：

- 航班取消的原因很难预测，有天气原因，乘客数量太少，甚至政治原因等，而且航班取消的比率不高以及由于航班取消，航空公司和保险公司无需提供延误险。
- 合并后的数据量较大，约为 700 万行，但是出现数据乱序以及缺失的比例不高，低于 2%。如图 4图 5图 6所示。

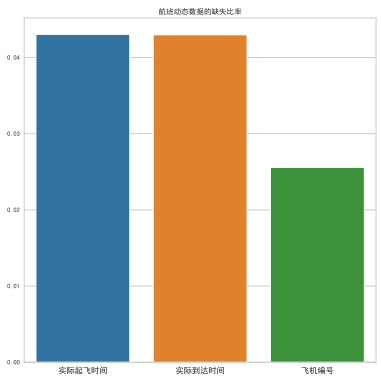

图 4. 航班动态数据的缺失比率

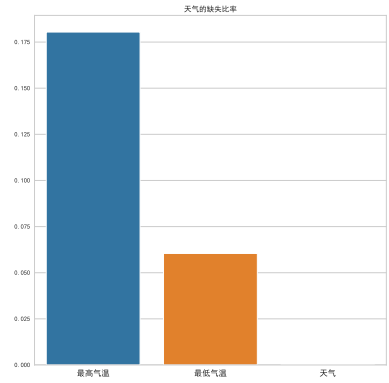

图 5. 天气的缺失比率

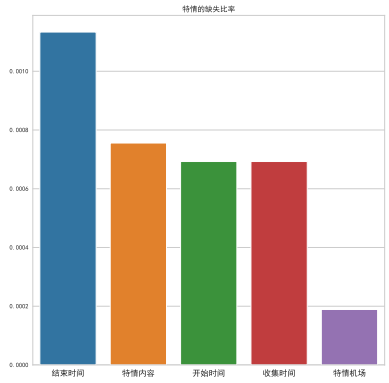

图 6. 特情的缺失比率

3.4 描述性分析

3.4.1 因变量

1. **飞机编号：**本数据集中不同编号的飞机延误概率从小于 0.02 到大于 0.06，飞机编号图 9可能对延误三个小时以上的概率有显著的影响，尤其是某些飞机延误的概率相对较高。这可能是飞机型号造成的。图 8为不同编号的飞机频率对比，而图 8为不同编号的飞机的平均延误概率对比，从这两个图中，我们可以看出，不同编号的飞机出现的频率以及航班延误概率有一定的差异。我们对出现频率比较高的飞机编号图 9作更仔细的分析。

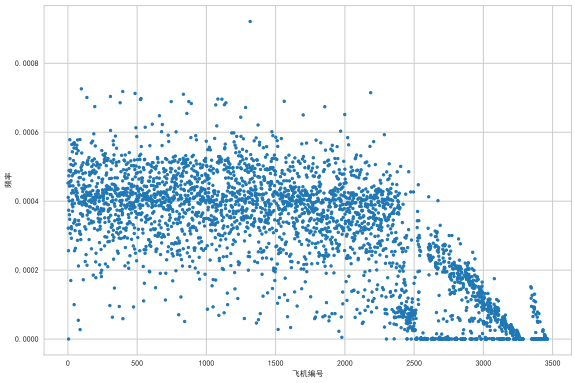

图 7. 不同飞机编号的频率比较

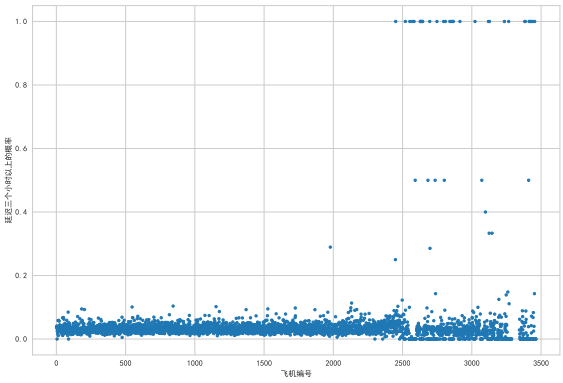

图 8. 飞机编号与航班延误关系

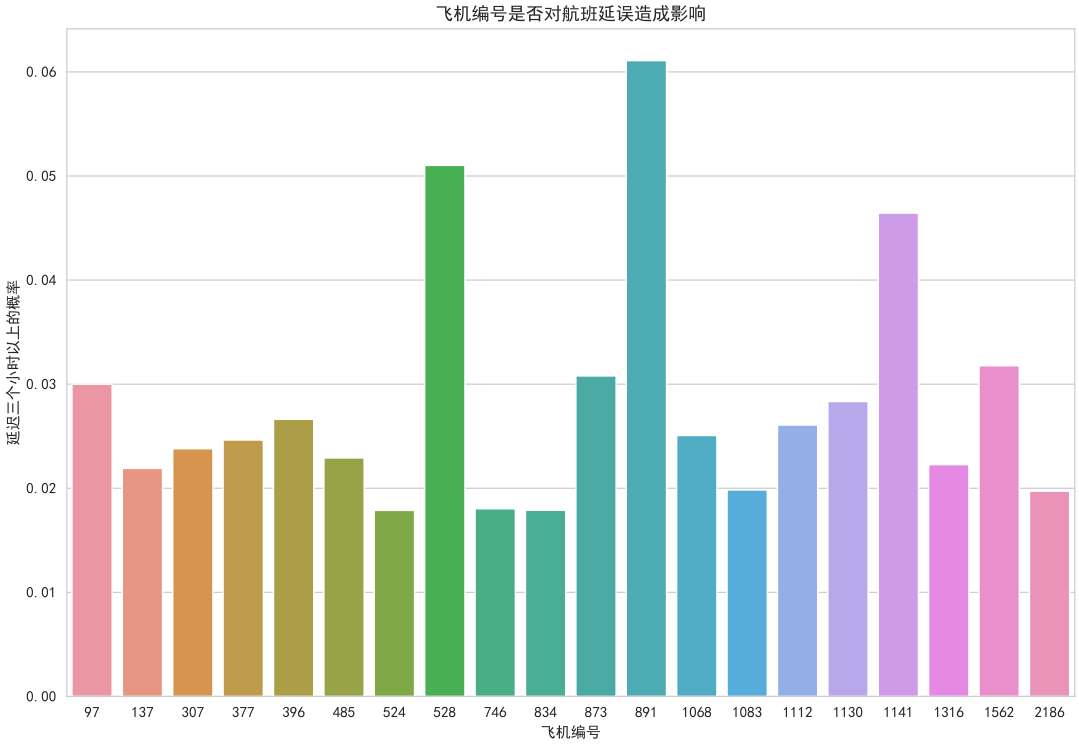

图 9. 飞机编号与航班延误关系，图中仅显示出现频率比较高的飞机编号的平均延误概率。

2. **机场：**图 10不同机场的延误概率从略高于 0.02 到大于 0.05，机场可能对延误三个小时以上的概率有显著的影响，尤其是某些机场延误的概率相对较高。这可能跟机场的基础设施和人流量有关。

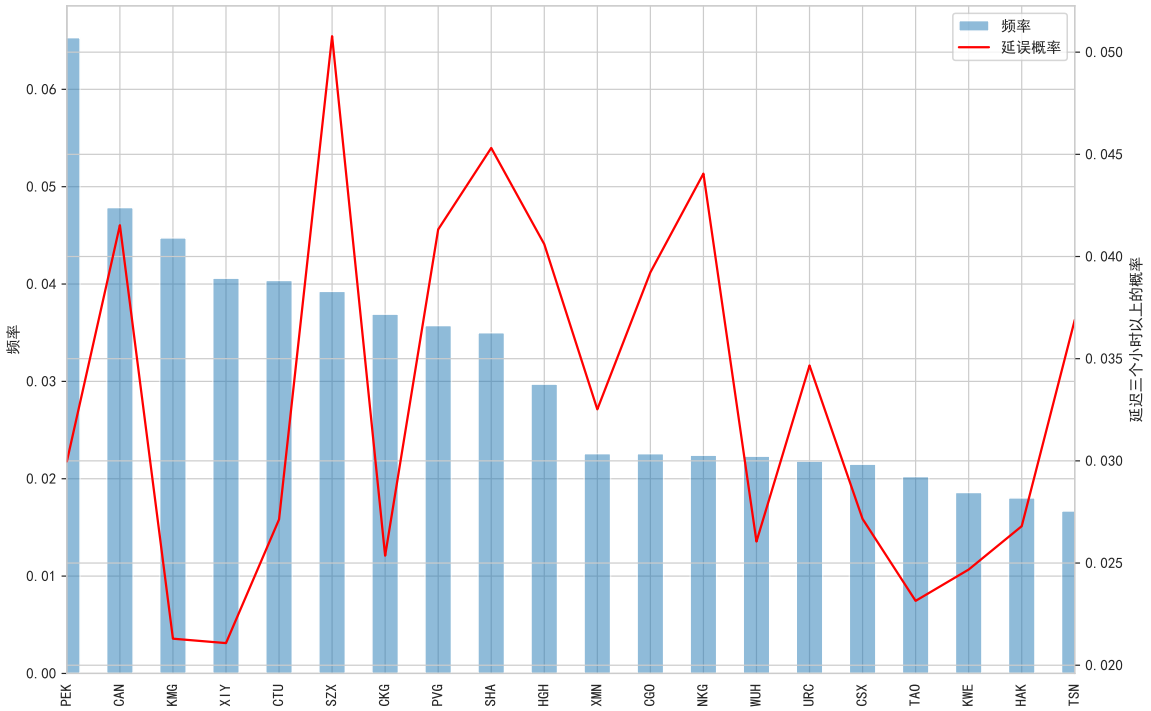

图 10. 机场与航班延误关系，图中仅显示出现频率比较高的机场的平均延误概率。

由于机场和城市存在相对应的关系，我们在中国地图上作出在不同机场（或城市）上起飞的飞机延误概率的热力图。如图 13所示。从图中可以看出，中国沿海地区（尤其是东南沿海地区）飞机延误概率较高，而内陆地区飞机延误概率较低。并且在大城市起飞的飞机的延误概率要比在小城市起飞的飞机的延误概率高。可能的原因是：

- 在大城市机场或者人口密集的城市机场，由于航班繁忙以及人流量较大，安全意识较强等，飞机延迟起飞的概率较高。如长三角地区机场，延误概率为全国最高。
- 在沿海城市，雷雨，大雨，大雪等极端天气出现的概率较高，会增加飞机延迟起飞的概率。如海南地区机场，人口密度甚至没有一些内陆城市高，但是由于全年降雨量大，恶劣天气较多，延误概率也较高。

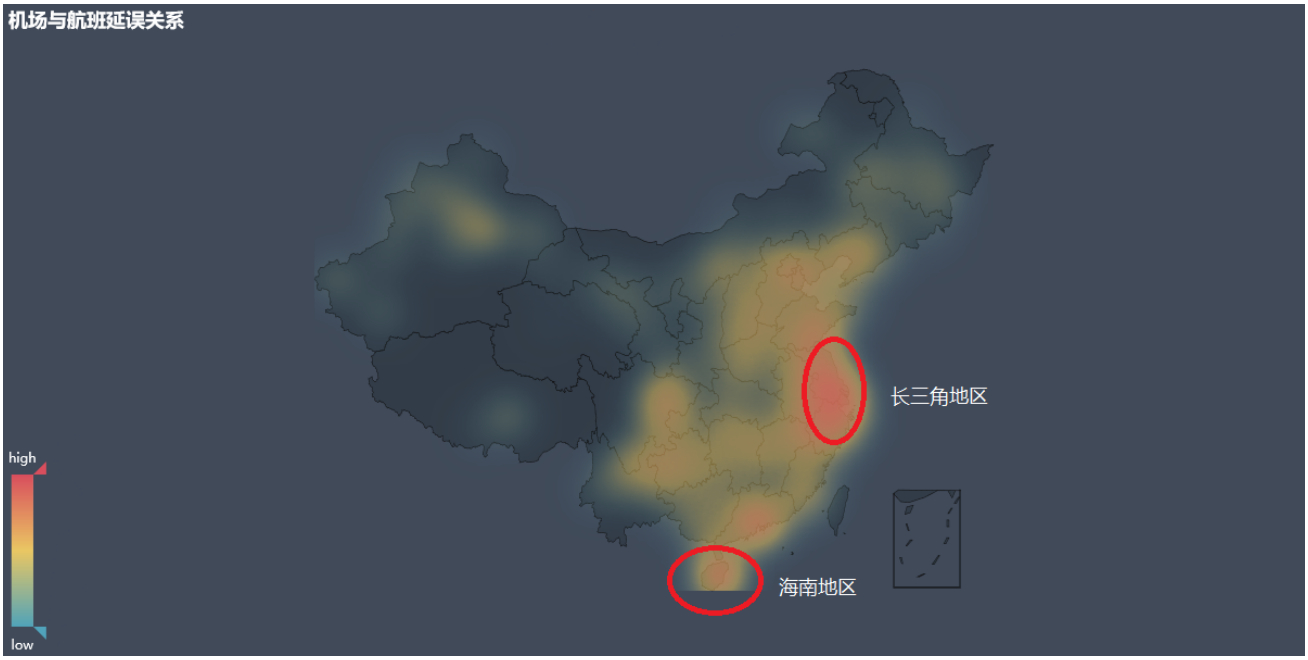

图 11. 在不同机场（或城市）上起飞的飞机延误概率的热力图，**长三角地区机场，**人口密集，**由于航班繁忙以及人流量较大，安全意识较强等，飞机延迟起飞的概率较高。**而**海南地区机场，**人口密度甚至没有一些内陆城市高，但是由于**全年降雨量大，恶劣天气较多，延误概率也较高。**

3. **航班编号：**不同航班图 12延误概率从小于 0.01 到大于 0.06，航班编号可能对延误三个小时以上的概率有显著的影响，可能是由于航空公司不同，而主要业务所在地区不同，而导致延误概率有显著差异，如图 23中的 KN 公司（中国联合航空有限公司）是一家主要业务在北京的公司，由可知道，在北京机场起飞的飞机的延误概率比较高。因此我们在图 23中探究了航空公司与航班延误的关系。另外，不同的航班编号可能经常使用不同类型的飞机，而不同类型的飞机都会对航班是否延误产生一定的影响。

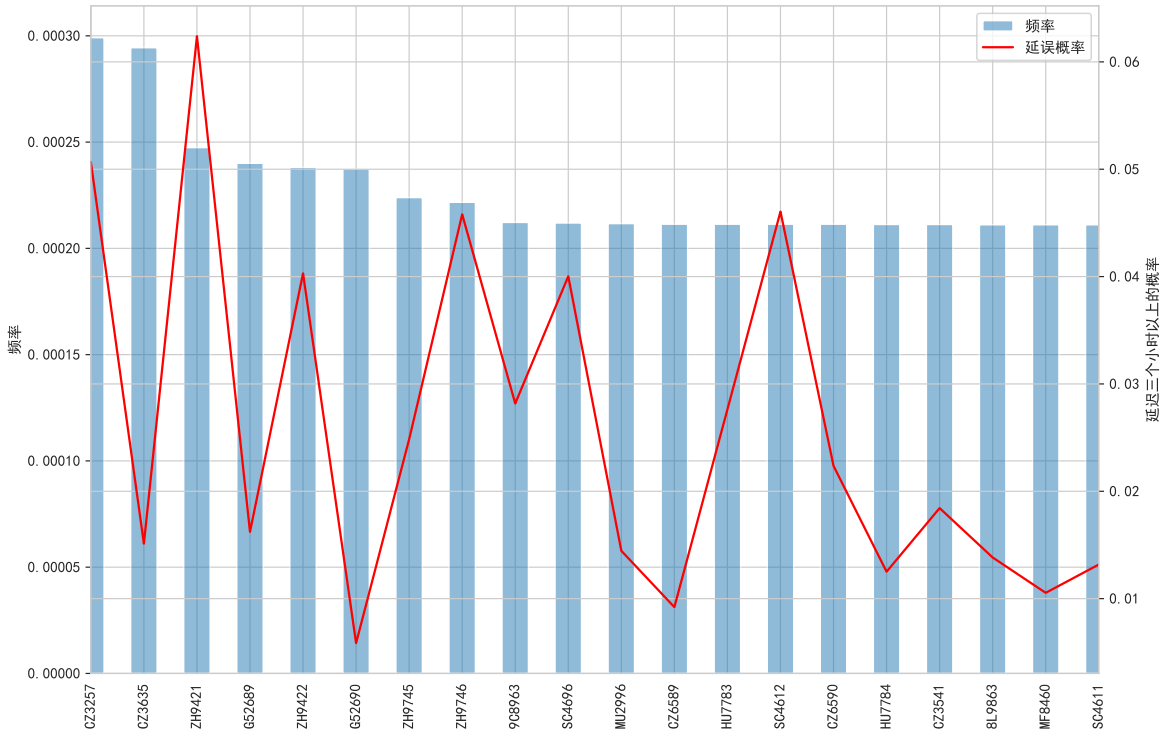

图 12. 航班编号与航班延误关系，图中仅显示出现频率比较高的航班编号的平均延误概率。

4. **航空公司：**不同航空公司的延误概率从 0.02 到大于 0.05 左右，这可能是因为航空公司的服务水平参差不齐。把航空公司作为一个重要变量，给乘客出行选择提供一个参考，也十分具有现实意义。

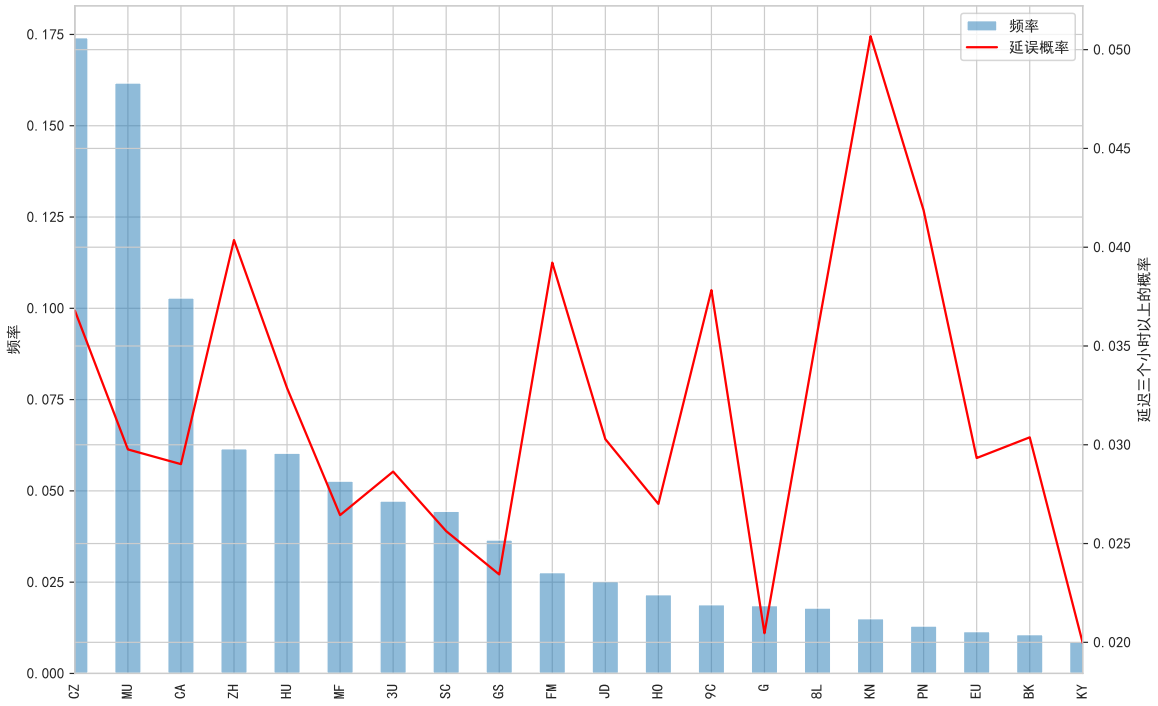

图 13. 航空公司与航班延误关系，图中仅显示出现频率比较高的航空公司的平均延误概率。KN 公司（中国联合航空有限公司）是一家主要业务在北京的公司，由可知道，在北京机场起飞的飞机的延误概率比较高。

5. **计划飞行时间：**用计划到达时间减去计划出发时间得到计划飞行时间图 14，从图 15可以看出计划飞行时间可能是一个重要的变量。

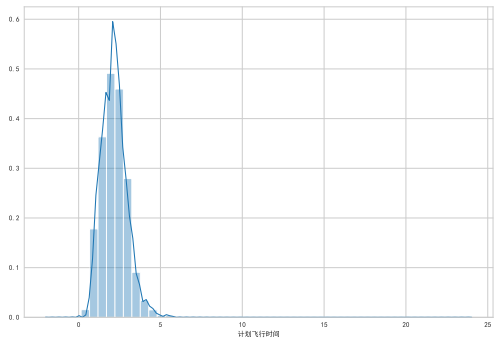

图 14. 计划飞行时间的分布。从图中可以看出，大部分航班的计划飞行时间在 1 ∼ 5 小时内，仅有少量超过五小时的航班（国际航班）。

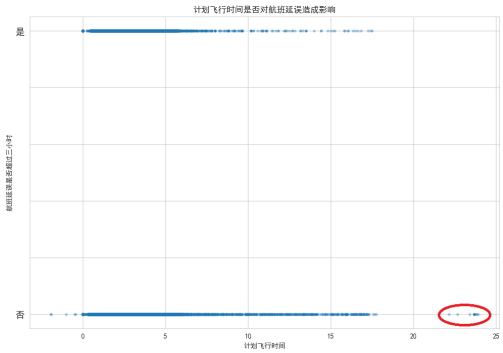

图 15. 计划飞行时间与航班延误关系。其中，横轴为计划飞行时间，纵轴为是否延误三小时以上。从图中可以看出，计划飞行时间越长，航班延误概率越高。但是**计划飞行时间超过一个阈值后，航班延误概率反而很低**，如红圈所示。

我们还对其进行了秩和检验，得到的统计量为 59.99，p-value < 0.001，因此我们认为航班是否延误跟计划飞行时间有一定的关系，表现为计划飞行时间越长，航班延误概率越高。

6. **天气**：天气对航班延误的概率从不到 0.02 到 0.1 左右，而极端天气影响的概率则是从 0.15 左右到接近 0.7 。从图中我们可以看出，天气是一个极其重要的变量，尤其是极端天气会大大提高飞机延误的概率。

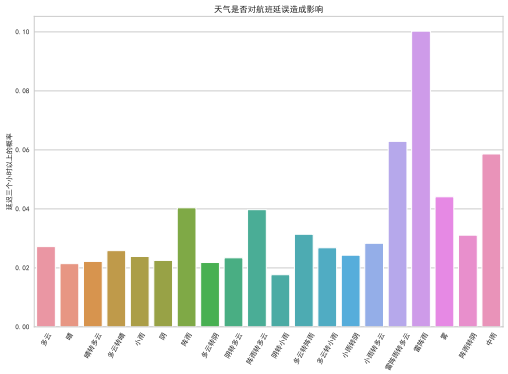

图 16. 天气与航班延误关系

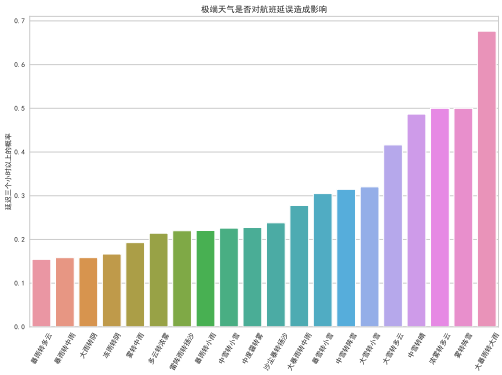

图 17. 极端天气与航班延误关系

7. **气温**：最低气温对航班延误的概率从小于 0.05 接近 0.4 不等，而平均气温的延误概率最高不到 0.3。但从日最高气温和日最低气温对航班延误的影响来看，极端气温会显著地提升航班延误的概率。

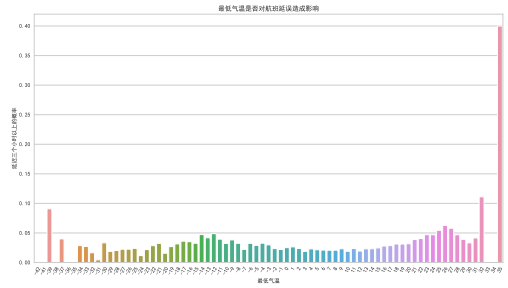

图 18. 最低气温与航班延误关系

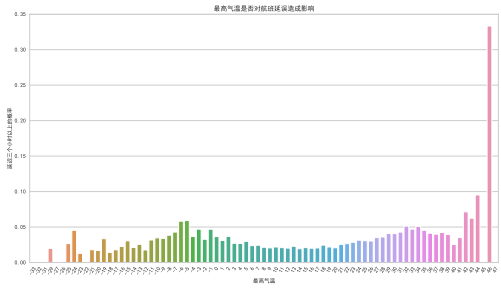

图 19. 最高气温与航班延误关系

8. **特情数量**：特情内容对航班延误的概率从小于 0.1 到接近 0,7 不等。条形图上的数字表示的是特情内容数量为该数值的航班出现的次数。从图中可以看出，随着特情数量的增加，航班延误概率有增加的趋势，但是也有一些特殊的情况，这可能是特情内容数量较多时，特情内容数量为该数值的航班出现的次数非常少，概率估计存在较大的误差。特情内容数量也是一个很重要的特征，尤其是特情内容数量较多时，航班延误的概率非常高。

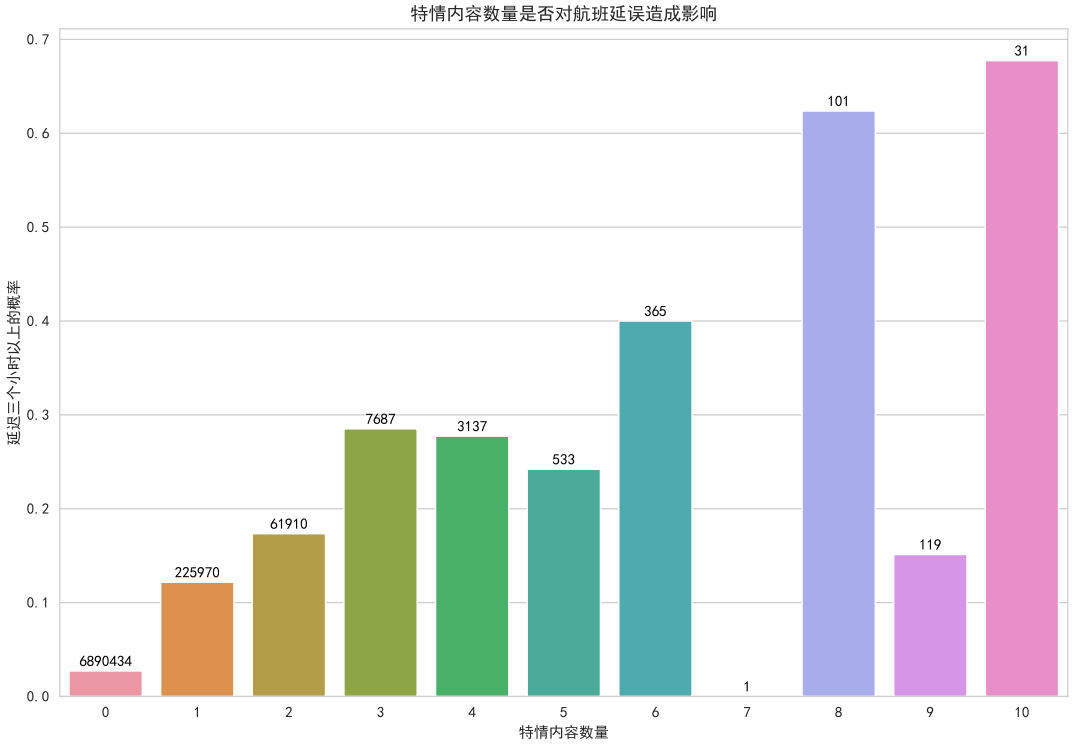

图 20. 特情数量与航班延误关系。条形图上的数字表示的是特情内容数量为该数值的航班出现的次数。

9. **特情内容**：对特情内容进行文本分析，不同的特情对航班延误影响概率从 0.05 左右到接近 0.3，选出对延误影响较大的关键词：雪，暴雨，能见度下降，能见度低。这也符合我们的常识。对不同的特情内容，我们可以计算特情内容紧急程度，计算步骤如下：

   - 提取特情内容中的关键词，如能见度下降，出现大面积延误等。
   - 计算含有这些特情内容的航班延误超过三个小时的概率。
   - 对于每个航班，计算其特情内容（如有）中含有的关键词对应的概率之和为该航班的特情内容紧急程度，没有特情内容则令特情内容紧急程度为 0。

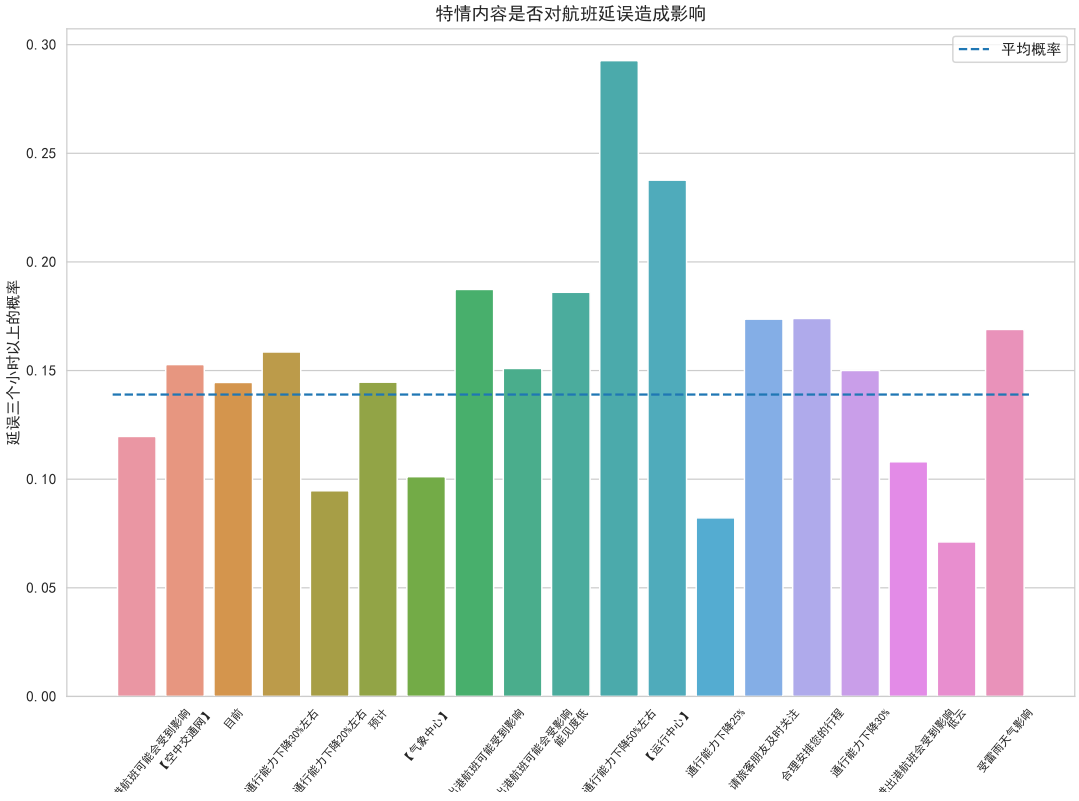

图 21. 特情内容与航班延误关系

10. **起飞时间**：起飞时间在凌晨的时候延误概率高达 0.2，而其余时间均低于 0.05，可见起飞时间也是一个非常重要的变量。从图 22图 23可以看到，凌晨的航班次数很少，并且延误概率非常高。

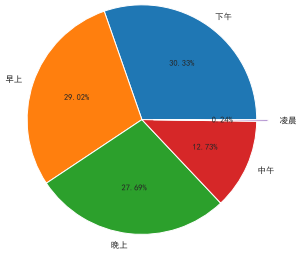

图 22. 不同起飞时间段的频率统计

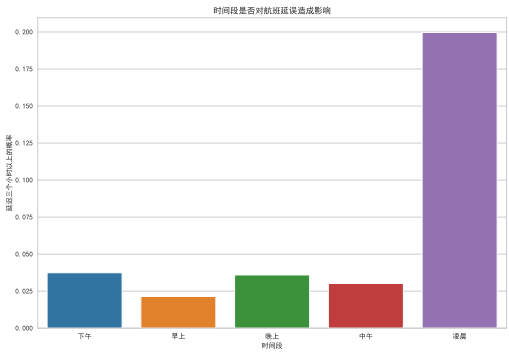

图 23. 起飞时间段与航班延误关系

11. **月份**：不同的月份，航空飞行的需求量不一样，比如在夏季和冬季，由于暑假，寒假和节假日的影响，航空需求量大大增加。需求量的增加会在一定程度上导致延误概率提升。另外，在夏季和冬季，恶劣天气出现的概率增加，也会导致延误概率提升。如图 25所示。

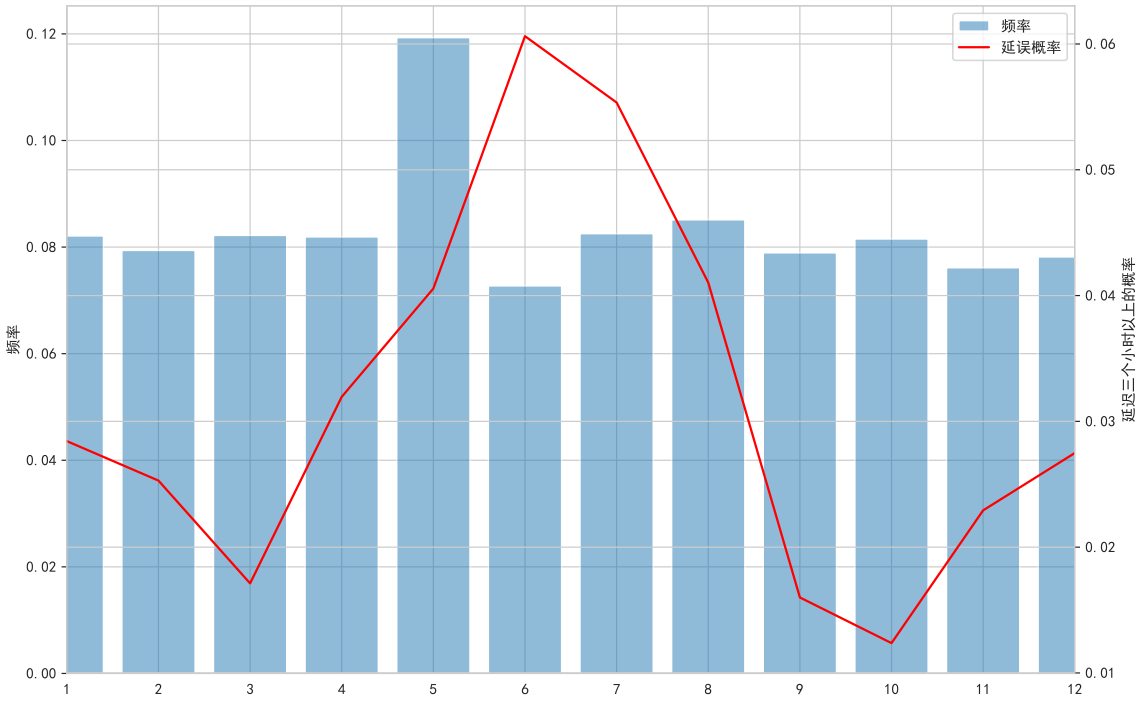

图 24. 月份与航班延误关系

经过特征工程，选出了 12 个重要的变量，分别是**飞机编号，航空公司，起飞时间段，机场，航班编号，计划飞行时间，天气，最高气温，最低气温，特情内容数量，特情内容和月份**。其中，飞机编号，机场，航班编号，天气，航空公司，月份和起飞时间段属于分类变量，计划飞行时间，平均气温和特情内容数量属于数值型变量，特情内容属于文本变量。

4 模型介绍

由于数据集较为庞大且分类种类较多，我们不能对分类变量使用 one-hot 编码，只能使用 label Encoder 编码（即编码为 1，2，...）。线性分类模型（如 Logistic 回归）对 encoder 编码不能起到很好的作用，因此我们选择非线性分类模型，如 Random Forest，XGBoost 等。

4.1 随机森林

作为新兴起的、高度灵活的一种机器学习算法，随机森林 [1,2]（Random Forest，简称 RF）拥有广泛的应用前景，从市场营销到医疗保健保险，既可以用来做市场营销模拟的建模，统计客户来源，保留和流失，也可用来预测疾病的风险和病患者的易感性。

1. 参数设定：

算法 1. 随机森林 [3] 模型算法

输入: 因变量 Y, 自变量$\boldsymbol{X}$, 树的个数 n_t

输出: $\{h(x,\theta^b)\}$

1: **当** $n \leq n_t$

2: 对$\boldsymbol{X}$ 进行 bootstrap 抽样得到$\boldsymbol{X^b}$

3: 从 r 维变量中选择 m 维变量 θ^m，利用选择的 m 个变量分裂节点，得到分类器 $h(x,\theta^b)$

4: **返回:** 组合 n_t 个回归树 $\{h(x,\theta^b)\}$

每次分裂选择的变量数为共变量的 1/3，每个叶子节点的最大样本量不超过 5，产生 300 棵树进行预测.

2. 变量重要性

4.2 Adaboost

Adaboost [4] 是一个累加的强分类器，能根据每次分类的结果进行权重的调节，表现为

$$F_T(x) = \sum_{t=1}^{T} f_t(x) \tag{1}$$

$f_t(x)$ 为对样本 x 进行分类的弱分类器。每个弱分类器都会对样本集中的每个元素进行预测，预测结果记为 $h(x_i)$，每次预测结束后会根据预测结果进行赋权，从而最小化预测误差 E_t, 如式 2

$$E_t = \sum_i E[F_{t-1}(x_i) + \alpha_t h(x_i)] \tag{2}$$

这里 $F_{t-1}(x_i)$ 是基于上一步分类结果构建的强分类器，$f_t(x) = \alpha_t h(x_i)$ 是在给定 t 步弱分类器权重的情况下，加到 t-1 步最终的分类器上。具体来说，在 t-1 步时我们强分类器 $C_{t-1}(x_i)$ 是弱分类器 $k_j(x_i)$ 的线性组合，形如式 3

$$C_{t-1}(x_i) = \alpha_1 k_1(x_i) + \cdots + \alpha_{t-1} k_{t-1}(x_i) \tag{3}$$

在第 t 步迭代我们继续强化我们的分类器如式 4所示

$$C_t(x_i) = C_{t-1}(x_i) + \alpha_t k_t(x_i) \tag{4}$$

我们利用 exponential loss 来定义我们的误差项式 5

$$E = \sum_{i=1}^{N} e^{-y_i C_t(x_i)} \tag{5}$$

对于 $m > 1$ 来说，令 $w_i^{(1)} = 1, w_i^{(t)} = e^{-y_i C_t(x_i)}$，我们有式 6

$$E = \sum_{i=1}^{N} w_i^{(t)} e^{-y_i C_t(x_i)} \tag{6}$$

我们可以重新整理求和，将分错和分对的样本分开如式 7

$$\begin{aligned} E &= \sum_{y_i=k_t(x_i)} w_i^{(t)} e^{-\alpha_t} + \sum_{y_i\neq k_t(x_i)} w_i^{(t)} e^{\alpha_t} \\ &= \sum_{i=1}^{N} w_i^{(t)} e^{-\alpha_t} + \sum_{y_i\neq k_t(x_i)} w_i^{(t)} (e^{\alpha_t} - e^{-\alpha_t}) \end{aligned} \tag{7}$$

为了求得 α_m 使得 E 最小，我们对 α_m 求导

$$\frac{dE}{d\alpha_t} = \frac{d(\sum_{y_i=k_t(x_i)} w_i^{(t)e^{-\alpha_t}} + \sum_{y_i\neq k_t(x_i)} w_i^{(m)} e_m^{\alpha})}{d\alpha_m} \tag{8}$$

将偏导令为 0 并解出 α_t 得式 9

$$\alpha_m = \frac{1}{2} \ln \left(\frac{\sum_{y_i=k_t(x_i)} w_i^{(m)}}{\sum_{y_i\neq k_t(x_i)} w_i^{(m)}} \right) \tag{9}$$

最后我们用求得系数 α_t 去更新 t 步的强分类器如式 10

$$C_m = C_{(m-1)} + \alpha_m k_m \tag{10}$$

4.3 GradientBoost

和 Adaboost 不同，GradientBoost [5–7] 在迭代的时候选择梯度下降的方向来保证最后的结果最好。损失函数用来描述模型的“靠谱”程度，假设模型没有过拟合，损失函数越大，模型的错误率越高。如果我们的模型能够让损失函数持续的下降，则说明我们的模型在不停的改进，而最好的方式就是让损失函数在其梯度方向上下降。

算法 2. GradientBoost 算法

$F_0(x) = \arg\min_\rho \sum_{i=1}^{N} L(y_i, \rho)$
当 $m \leq M$
  $\tilde{y}_i = -\left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}$
  $\alpha_m = \arg\min_{\alpha,\beta} \sum_{i=1} N[\tilde{y}_i - \beta h(x_i;\alpha)]^2$
  $\rho_m = \arg\min_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i;\alpha_m))$
  $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x};\alpha_m)$

这里是直接对 Adaboost 模型的函数进行更新，利用了参数可加性推广到函数空间。训练 $F_0 - F_m$ 一共 m 个基学习器，沿着梯度下降的方向不断更新 ρ_m 和 α_m。

4.4 XGboosting

XGBoost 是以 CART 树中的回归树作为基分类器，在给定训练数据后，基本确定其单个树的结构 (叶子节点个数、树深度等等)。但 XGBoost 并不是简单重复的将几个 CART 树进行组合。它是一种加法模型，将模型上次预测 (由 t-1 棵树组合而成的模型) 产生的误差作为参考进行下一棵树 (第 t 棵树) 的建立。以此，每加入一棵树，将其损失函数不断降低。

XGboosting 的预测模型可表示为

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

其中 K 为树的总个数,f_k 表示第 k 颗树,$\hat{y}_i$ 表示样本 x_i 的预测结果。

损失函数表示为

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{K} \Omega(f_k)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

$$f_k(x) = w_{q(x)}, W \in R^T, q : R^d \leftarrow \{1, 2, \ldots, T\}$$

其中 $l(y_i, \hat{y}_i)$ 为样本 x_i 的训练误差,$\Omega(f_k)$ 表示为第 k 棵树的正则项,T 为树的个数,w 为叶子节点得分值。

将目标函数二阶泰勒展开

$$\begin{aligned}
Obj^t(\theta) &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{K} \Omega(f_k) \\
&= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\
&\approx \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) f_t(x_i) + \frac{1}{2}\partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) f_t(x_i)^2] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\
&= \sum_{i=1}^{n} [g_i w_{q(x_i)} + \frac{1}{2} h_i w^2_{q(x_i)}] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} + C \\
&= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \\
&= \sum_{j=1}^{T} [G_j w_i + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T
\end{aligned}$$

求导可得 $w_j^* = -\frac{G_j}{H_j + \lambda}$, 得到最终的目标函数

$$Obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$

其中

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$$

$$G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i$$

γ, λ 为模型的惩罚项。

综上，可利用最终求得的目标函数 Obj^* 进行模型求解,

4.5 LightGBM

LightGBM 不同于 XGBoost 是直接去选择获得最大收益的结点来展开，而 XGBoost 是通过按层增长的方式。以这种方式 LightGBM 能够在更小的计算代价上建立我们需要的决策树。在算法中我们还可以控制树的深度和每个叶子结点的最小数据量，从而减少过拟合。

概括来说，lightGBM 主要有以下特点：

1. 基于 Histogram 的决策树算法
2. 带深度限制的 Leaf-wise 的叶子生长策略
3. 直方图做差加速
4. 直接支持类别特征 (Categorical Feature)
5. Cache 命中率优化
6. 基于直方图的稀疏特征优化
7. 多线程优化

5 模型实现以及模型评价

通过特征工程我们找到了飞机延误的主要因素,开发了飞机延误预测模型。共使用了包括 LightGBM,
Random Forest 等的五种非线性分类模型预测飞机是否延误，对训练集使用了 5-fold 的交叉验证。

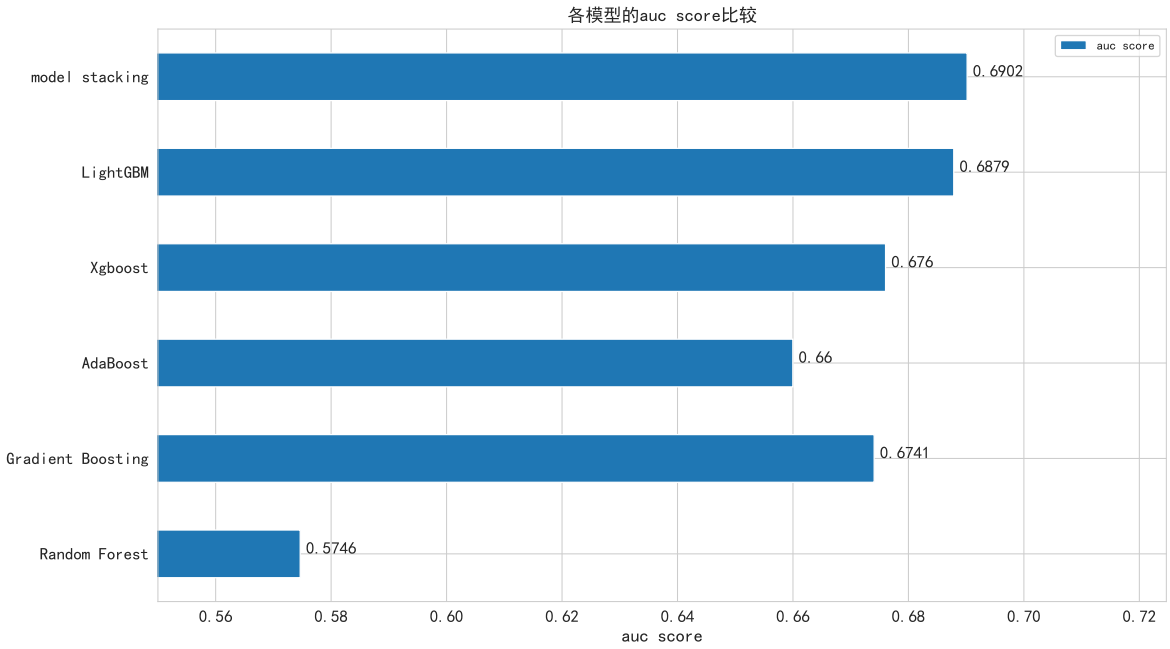

图 25. 各模型的 AUC 比较

从图 25可以看出，各个模型在 5-fold 数据集上的 AUC 如上图所示，LightGBM 模型在不同数据集上的 AUC 都是最高的，所以 LightGBM 模型的泛化能力明显优于其他四种模型。图 26是各个模型的交叉检验得到的平均 ROC 曲线比较图：

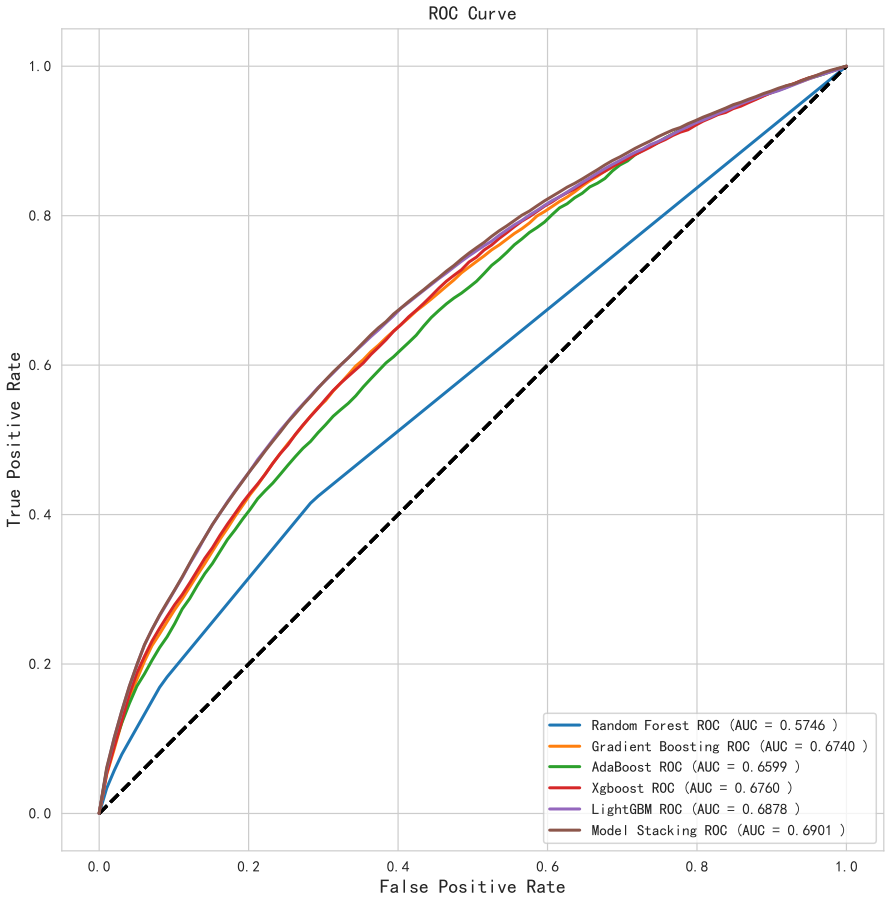

图 26. 各模型 ROC 曲线

我们以 8:1:1 和时间顺序划分训练集，验证集和测试集，我们在训练集上训练模型，分别在验证集和测试集上计算预测结果（延误概率），上图表示各模型在测试集上，从图中可知，五种模型的 ROC 曲线均在 AUC = 0.5（图中黑色虚线）的上方，所以各模型的 AUC 均是大于 0.5 的，各模型对飞机是否延误均具有一定的预测能力。ROC 曲线下的面积最大的是 LightGBM 模型的（即图中橙色曲线），AUC=0.6879，预测准确度最高。比较这五种模型，可以看到 LightGBM 模型在泛化能力和预测的准确性上都优于其他四种模型。

我们对五个分类模型进行融合，**融合后模型的 AUC 可以达到 0.6902，超过了在比赛中获得第一名的成绩 0.60**。模型融合的步骤如下所示。

- 以五个模型在验证集上的预测结果为 predictor，以验证集上真实的标签作为 target，进行 Lasso 回归，并且使用五折交叉验证选择最佳的惩罚系数 α。
- 以上面训练的 Lasso 模型预测测试集上的 target（以五个模型在测试集上的预测结果为 predictor）。

- 对回归系数进行微调。

6 总结

通过以上结果，我们对乘飞机出行的旅客提出以下几点建议。

1. 购票时注意飞机编号。因为不同的编号对应不同类型的飞机（比如大飞机和小飞机），大飞机可能更不容易延误；
2. 尽量选择距离自己最近的大型机场的航班，如首都国际机场，避免购买和乘坐小型机场的飞机；
3. 留意航班编号。航班编号就像人的身份证一样，如可以从编号中看到该飞机隶属于哪家航空公司。应尽量选择乘坐大型的国际航空公司的飞机；
4. 规划好自己的飞行时间。不同的飞行时间对航班延误也是有一定影响的；
5. 查看出行近期的天气和气温。恶劣天气（如台风，大雪等）容易造成航班延误甚至取消航班，所以应该选择好天气乘坐飞机出行。以上仅对于外出旅游等可以自由规划时间的旅客提的建议，对于因不可抗力等因素必须出行的旅客来说，就不必考虑以上那么多的因素，只要选择就近机场，留意天气和其他自身因素即可。

7 后续阶段检验

后续阶段，我们将该模型用于实际的航班延误预测，并且计算预测准确率，召回率，F1 score 等，以检验该模型的可靠性以及提供进一步改进模型的依据。

参考文献

[1] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[2] Alan Julian Izenman. Modern multivariate statistical techniques. *Regression, classification and manifold learning*, 2008.

[3] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[4] Gunnar Rätsch, Takashi Onoda, and K-R Müller. Soft margins for adaboost. *Machine learning*, 42(3):287–320, 2001.

[5] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[6] Ping Li, Qiang Wu, and Christopher J Burges. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in neural information processing systems*, pages 897–904, 2008.

[7] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 2013.