第一届太普华南杯数据挖掘建模竞赛

题　　目　**关于航空客户的信息挖掘**

队　　长　**程裕**

成　　员　**蔡少真　　李伟东**

学校(院系)　**华南理工大学（理学院数学系）**

指导教师

完成时间　**2013-4-21**

综合评定成绩：_____

评委评语：

评委签名：

关于航空客户的信息挖掘

摘　要： 为了提高航空的上座率，对样本数据进行挖掘进行客户流失预测、客户细分及客户价值评估。基于 logistic 回归分析建立客户流失预测模型，得出每个客户的流失倾向概率。定义一阈值为 0.5，若流失倾向概率大于 0.5，则该客户的预测状态为流失；反之，则非流失。 建立 RFM 模型将客户划分为重要保持客户、重要发展客户、重要挽留客户、忠诚型一般客户、低价值客户五个类别。最后，综合分析客户的类型和流失状态，分析不同客户的特征，得到以下结果：

- 不同类别客户的指标平均值都不相近，易识别；
- 单从各个类别客户来讲，各指标情况在流失客户与非流失客户中有一定的差异。

针对不同的客户，可以采取不同的营销活动来提高上座率：

- 针对重要保持客户，进行客户保持。
- 针对重要发展客户，进行客户发展。
- 针对重要挽留客户，进行流失预警。
- 针对忠诚型一般客户，进行交叉销售。

关键词：客户流失预测　客户细分　　客户价值评估　　logistic 回归分析　RFM 模型

Data Mining on Aviation Customers

Abstract: In order to increase the attendance rates on airlines, mining customer churn prediction was carried out on the sample data, customer segmentation and customer value evaluation. Based on logistic regression analysis to establish customer churn prediction model, each customer tendency of loss probability are obtained. To define a threshold value 0.5, if the loss probability is greater than 0.5, then we define that customer to be lost; on the other hand, not lost. Establish RFM model to help deviding the customers into 5 classes:important maintaining customers, important developing customers, important retaining customers, loyal customers, low value customers.At last,considering the types and loss status of customers,synthetically analyse the characteristics of different customers.The results are:

- different kinds of customers' indicators are different,easy to identify
- lost and not lost customers' indicators are still different

According to different customers, different marketing activities can be taken to increase the attendance of the airlines:

- For important maintaining customers:keep good relationship with these customers
- For important developing customers:develope positive relationship with these customers
- For important retaining customers:prevent customer churn
- For loyal customers:conduct crossed marketing

Key words: customer churn prediction customer segmentation customer value elauation logistic regresson analysis RFM model

目 录

目录

1. 挖掘目标

市场竞争异常激烈的今天，如何识别有价值的客户是企业营销策略的一个非常重要的环节。我们希望通过从大量的旅客乘机记录中对航空公司的客户进行行为分析，采用数据挖掘技术，达到以下目标：

- 对客户进行流失倾向评分，预测流失情况；
- 进行客户细分，将客户划分为五类；
- 客户价值评估，挖掘出有价值的客户；
- 综合分析客户流失与客户细分结果，提出有效方案以进行更精确地营销，从而实现提升航空客运的上座率目标。

2. 分析方法与过程

2.1. 客户流失模型

客户流失是指客户因某种原因而离开为其服务公司的一种常见行为。由于各种因素的不确定性和市场不断的增长以及一些竞争对手的存在，很多客户不断地从一个公司转向另一个公司，其目的是为了求得更低的价格和更好的服务。一般来说，流失客户可分为自愿流失和非自愿流失，而航空公司的流失客户基本上是属于自愿流失的。

客户流失预测主要是对客户现所处状态的一种预测，通过模型计算出客户流失倾向概率，给定一阈值与概率进行比较。当流失倾向概率大于阈值时，则将该客户预测为流失；若流失倾向概率小于或等于阈值时，则预测结果为非流失。

在本题中，我们定义流失客户为：最后一次乘机时间至观察窗口末端时长>=观察窗口内最大乘机间隔。并以 1 标记流失客户，0 标记为非流失客户。总共可得到非流失客户数为 38519，流失客户为 24468。

2.1.1. 数据样本

同时尽可能收集能影响客户流失的各种因素，包括：入会时间，第一次飞行时间、性别、会员卡级别、年龄、飞行次数、基本积分、总加权飞行公里数、平均乘机时间间隔、其他积分、非乘机的积分变动次数等等。

为了能更好的分析数据随季度的变化情况，我们引入了趋势值和变动值：

趋势值：表示 8 个季度内属性增大或减小的速度与方向，以一元线性回归的斜率表示。斜率大于 0，表示增大，斜率越大，增加速度越大；斜率小于 0，表示减小，且斜率越小，减小的速度越大。

波动值：表示 8 个季度内属性的变化幅度，以样本的方差表示。方差越大，表示数据变化幅度越大越不稳定；方差越小，表示数据变化幅度越小越稳定。

对于数据缺失的情况，SPSS 中带有处理缺失值的方法：

①剔除法

当缺失值非常少的时候，可对缺失的数据进行删除或报告。

②替代法

SPSS 中可以选择以变量均值、临近点的均值、临近点的中位值、线性内插发或线性趋势法来替换缺失的数据。

2.1.2. **Logistic 回归分析**

客户流失状态只有两种情况，即流失与非流失，这两种状态分别用 1 和 0 表示。因此我们可以采用多因素非条件 logistic 回归模型为基本依据，通过 logistic 回归建立客户流失概率预测模型进行评价，从而得出每个客户的流失倾向概率。

设客户的流失情况为

$$Y_i = \beta_0 + \sum_1^p \beta_j x_{ij} + \varepsilon_i \quad (1)$$

$Y_i = 1$表示流失，$Y_i = 0$表示非流失，$\varepsilon \sim N(0, \sigma^2)$，$E(\varepsilon) = 0$，则有

$$E(Y_i) = \beta_0 + \sum_1^p \beta_j x_{ij} = 1 * P + 0 * (1 - P) = P \quad (2)$$

即$E(Y_i)$为x_i时$Y_i = 1$的概率值。

由于概率 P 的取值范围是在[0,1]区间，需先对概率 P 做 Logit 变换，具体如下：

第一步，将 P 转换成Ω，即

$$\Omega = \frac{P}{1-P} \quad (3)$$

称Ω为发生比，是事件发生的概率与不发生的概率的比值。

第二步，将Ω转换成 lnΩ，即

$$\ln\Omega = \ln\left(\frac{P}{1-P}\right) \quad (4)$$

称lnΩ为 Logit P，经过变化后的Ω与 Logit P 之间的增长性是一致的。

经过 Logit 变化后，则可建立自变量和因变量之间的关系模型，即逻辑回归模型：

$$\text{Logit}\,P = \beta_0 + \sum_1^p \beta_j x_{ij} \quad (5)$$

即

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_1^p \beta_j x_{ij} \quad (6)$$

于是有

$$\frac{P}{1-P} = \exp(\beta_0 + \sum_1^p \beta_j x_{ij}) \quad (7)$$

从而有

$$P = \frac{1}{1+\exp[-(\beta_0+\sum_1^p \beta_j x_{ij})]} \quad (8)$$

即为逻辑回归函数，是 典型的增长函数，能很好体现概率 P 和自变量间的非线性关系。

得到每个客户的流失倾向概率后，给定阈值 0.5，当流失倾向概率大于 0.5，则预测结果为流失；若流失倾向概率小于 0.5，则预测结果为非流失。

2.1.3. **结果分析**

借助 SPSS 软件对数据进行 logistic 回归分析，筛选出对模型影响较大的指标，经过多次筛选结果显示观测窗口季度平均飞行次数 X1、积分兑换次数 X2、非乘机的积分变动次数 X3、平均乘机时间间隔 X4、飞行次数波动值 X5、飞行次数趋势值 X6、年龄 X7 对回归模型较为显著影响。

Logistic 回归分析结果为：

表一　Logistic 回归分析模型系数综合检验

模型系数的综合检验

		卡方	df	Sig.
步骤 1	步骤	38125.099	7	.000
	块	38125.099	7	.000
	模型	38125.099	7	.000

表二 Logistic 回归分析模型汇总

模型汇总

步骤	-2 对数似然值	Cox & Snell R 方	Nagelkerke R 方
1	46109.987	.454	.616

表三 Logistic 回归分析分类表

分类表 a

已观测			已预测		
			是否流失		百分比校正
			0	1	
步骤 1	是否流失	0	35283	3150	91.8
		1	5033	19521	79.5
	总计百分比				87.0

表四 Logistic 回归分析方差中的变量

方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)
步骤 1a	观测窗口季度平均飞行次数	-2.483	.028	7963.733	1	.000	.083
	积分兑换次数	.048	.018	6.972	1	.008	1.049
	非乘机的积分变动次数	-.010	.002	24.602	1	.000	.990
	平均乘机时间间隔	-.027	.000	7310.059	1	.000	.974
	飞行次数波动值	.151	.009	303.160	1	.000	1.164
	飞行次数趋势值	-4.542	.055	6713.011	1	.000	.011
	年龄	-.006	.001	28.043	1	.000	.994
	常量	3.852	.062	3830.930	1	.000	47.068

结果显示，模型的卡方值较大，Sig=0.000，说明模型整体是显著的。而 Cox & Snell R 方 Nagelkerke R 方的值在 0.5 左右，说明模型拟合情况一般。预测的准确值达 87%，这情况还是令人满意的。同时 7 个指标的 sig<0.01，即自变量与 Logit P 之间的线性关系是显著的。

利用逐步 logistic 回归分析法建立的模型，建立客户流失预测模型：

$$p(y) = \frac{\exp(-2.483x1 + 0.048x2 - 0.01x3 - 0.027x4 + 0.151x5 - 4.542x6 - 0.006x7 + 3.852)}{1 + \exp(-2.483x1 + 0.048x2 - 0.01x3 - 0.027x4 + 0.151x5 - 4.542x6 - 0.006x7 + 3.852)}$$

以 p(y)为应变量（客户流失的预测概率），采用 ROC 曲线评价 logistic 回归分析预测模型，结果为：

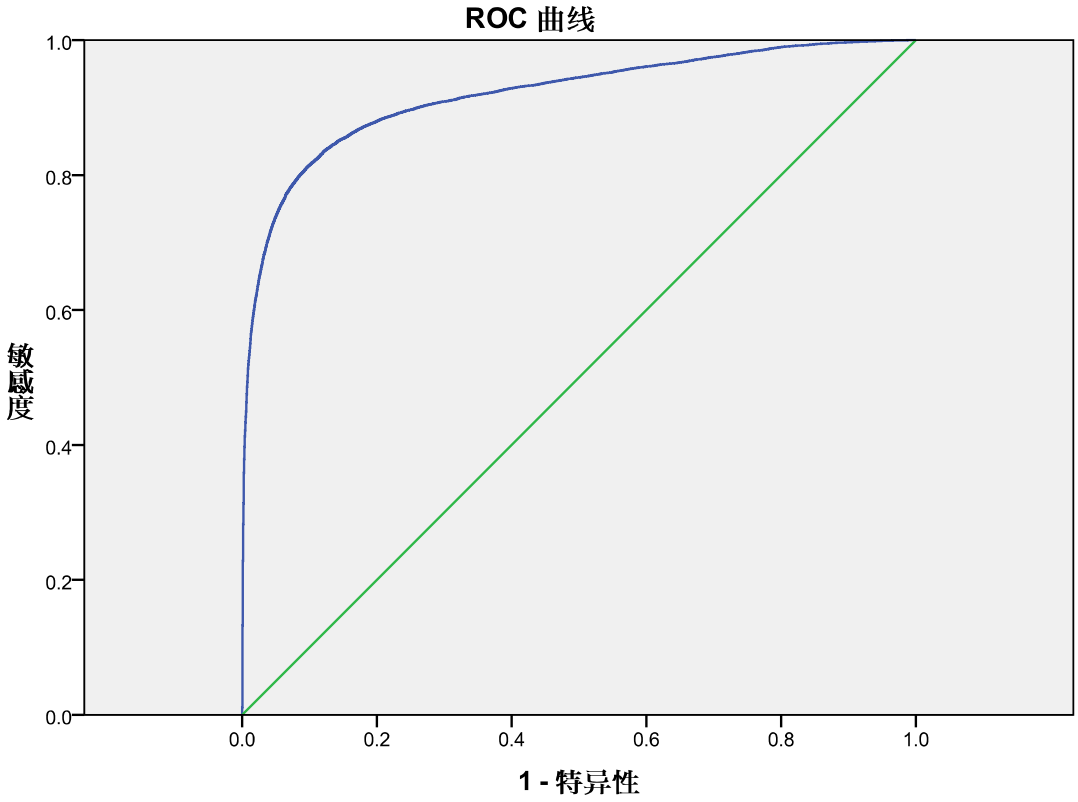

图 1　ROC 曲线图

曲线下的面积

检验结果变量:预测概率

面积	标准误[a]	渐进 Sig.[b]	渐近 95% 置信区间	
			下限	上限
.919	.001	.000	.917	.922

检验结果变量:预测概率 在正的和负的实际状态组之间至少有一个结。统计量可能会出现偏差。

a. 在非参数假设下

b. 零假设：实面积 = 0.5

ROC 曲线下面积为 0.919，与机会线下的面积比较具有显著性差异，Sig=0.000<0.01，说明所建立的概率预测模型预测效果显著，能较为准确的预测客户流失的概率。

（客户流失模型所需数据及预测结果详见“附件 1 客户流失预测模型.xls”）

2.2. 客户细分及客户价值评估

客户细分的概念是美国市场学家温德尔.史密斯于 20 世纪 50 年代中期提出来的。所谓客户细分，是以消费者需求为出发点，根据消费者购买行为的差异性，把消费者总体划分为类似性购买群体的过程。分属于同一客户群的消费者具备一定程度的相似性，而不同的细分客户群间存在明显的差异性。

客户细分的目的，就是要更精确地回答谁是我们的客户，客户到底有哪些实际需要，企业应该去吸引哪些客户，应该重点保持哪些客户，应该如何迎合重点客户的需求等重要问题，进而使客户关系管理真正成为业务获得成功、扩大产品销量的助推器。客户细分是客户关系管理的基础，也是核心。根据帕累托定律（2/8 定律）：20%的顾客给企业带来 80% 的销售利润。企业投入大量资源来争取客户的目标应该定位于能为企业带来大量利润的那小部分客户群体，让他们长期成为自己的关系客户，而由此可以节省争取其他客户的成本。

航空公司的各种资源是有限的，面对数量众多的常旅客，如何对这些客户进行细分，更有效地判断有价值客户，了解他们的特征和实际需求，对不同价值的客户采取不同的营销策略，将有限的资源投放到最有价值的客户身上，实现精准化营销，提高企业的竞争力，最终实现提升航空客运的上座率目标。

2.2.1. 航空公司客户细分参数的确定

基于许多数据库营销的经验，营销专家鲍比斯通(Bob Stone)提出了 RFM 细分，通过三项变量，即最近消费时间间隔(Recency)、消费频率(Frequency)、消费金额(Monetary)来细分客户，用于识别最有价值的客户。RFM 指标高的客户可能更愿意并更有兴趣与企业进行交易，因此具备更高的客户价值。RFM 指标较差的客户代表了较少的业务机会，表明该客户价值较低。因此，RFM 模型可以帮助企业决定向谁促销，采用最优化的营销手段来获取和保留最有价值的客户，避免投资到效果很差的客户身上，吸引高价值客户，并以此来增进客户忠诚。

研究证明，基于权重的 RFM 方法是一种有效的客户细分方法。我们根据航空公司独有的特点，对传统的 RFM 指标做了适当的调整，确定了 L、R、F、M、C 五个指标作为航空公司客户细分的参数。L 代表客户关系长度（从入会之日算起），R 代表客户最近一次消费距今时间长度，F 代表客户在一定时间内的消费频率，M 代表客户在一定时间内的升级里程（“升级里程”是指航空公司会员通过乘坐本公司及航空合作伙伴的有效航班所累积的基本飞行里程），C 代表客户在一定时间内所乘航班的平均舱位折扣系数（即会员在一定时间内乘坐舱位对应折扣系数平均值）。（本文章取一定时间为八个季度）

2.2.2. 航空公司客户细分的具体步骤

①主要分析方法

(1)对于五个指标的重要程度不同，在查阅资料后，我们给定 L、R、F、M、C 的相对重要性（即权重）如下：

L=0.1，R=0.1，F=0.2，M=0.2，C=0.4

（2）快速聚类法。快速聚类法计算量非常小，从而可以有效地处理多变量大样本数据而不占用太多的内存空间和计算时间。

②基本步骤

（1）将 LRFMC 各指标标准化（反向指标 R 要与正向指标区别处理）并加权；

（2）确定聚类的类别数量为 5，应用快速聚类法对加权后的指标进行聚类；

（3）将每类客户的 LRFMC 平均值和总 LRFMC 平均值作比较，通过对比得到每类客户 LRFMC 的变动情况。分析每类客户的特点，在此基础上定义客户类型；

（4）计算每类客户的客户价值总得分，可将航空公司的客户群体划分成重要保持客户、重要发展客户、重要挽留客户、忠诚型一般客户、低价值客户等五个级别。针对不同等级的客户，航空公司可以采取不同的管理策略。

（5）客户价值比较分析。客户分类后，并不知道每一类客户的价值差别有多大，相对企业的重要性怎样。根据比较每类中心各指标的平均值加权得分的大小来对各类客户进行排序，客户价值排名靠前的客户相对排名靠后的客户对于企业来说更为重要。

2.2.3. **结果分析**

根据上述步骤，在 SPSS 软件中进行操作。结果显示，经过 100 次迭代后，聚类中心内没有改动或改动较小而达到收敛。聚类情况如下图所示。（用于聚类分析的数据详见“附件 2 聚类分析初始数据”，分析所的数据详见详 “附件 3 聚类分析后各客户对应类别.xls”）

表五 客户细分聚类的个数

每个聚类中的案例数

聚类	1	11884.000
	2	2364.000
	3	22656.000
	4	1783.000
	5	24301.000
有效		62988.000
缺失		.000

再对各聚类中心指标进行加权计算。，得到各类加权得分如下：

表六 各类客户的加权得分

1	0.073386929	重要挽留客户
2	0.520864142	重要保持客户
3	-0.179564518	低价值客户
4	0.424153722	重要发展客户
5	0.049730239	忠诚型一般客户

我们即可针对不同类别的客户进行不同的营销策略，提高航空公司的上座率。

3. 结论

本文通过对航空公司的会员数据进行客户流失模型、客户细分和客户价值评估分析，将由客户流失预测模型得出的客户流失状态和 RFM 模型得出的客户类别进行综合分析。对于航空公司在提高上座率这方面来说，更需要的是应针对各类客户不同情况能够采取恰当行动的完整方案，而非仅仅一个名单。

分别统计五类客户的不同流失状态下会员卡级别、年龄、观测窗口最大乘机间隔、总累计积分、观测窗口总加权飞行公里数的情况：

第一类：重要挽留客户

	会员卡级别	年龄	观测窗口内最大乘机间隔	总累计积分	观测窗口总加权飞行公里数
否	4.192774194	43.6804097	136.4117419	23929.42581	25601.17923
是	4.191340106	42.86598689	135.9242864	23942.49129	25522.96354

第二类：重要保持客户

	会员卡级别	年龄	观测窗口内最大乘机间隔	总累计积分	观测窗口总加权飞行公里数
否	4.47627965	49.36133923	164.9082397	25852.00499	24728.39703
是	4.49343832	48.11523179	168.6614173	24193.23622	23162.11996

第三类：低价值客户

	会员卡级别	年龄	观测窗口内最大乘机间隔	总累计积分	观测窗口总加权飞行公里数
否	4.010704979	41.86975624	177.1312064	5190.071132	5787.092592
是	4.008158922	40.91885782	175.9276339	4864.015963	5533.286255

第四类：重要发展客户

	会员卡级别	年龄	观测窗口内最大乘机间隔	总累计积分	观测窗口总加权飞行公里数
否	5.224283305	44.43776461	70.59106239	87180.74958	79384.16748
是	5.206030151	44.48653199	72.18257956	90929.74372	80912.77682

第五类：忠诚型一般客户

	会员卡级别	年龄	观测窗口内最大乘机间隔	总累计积分	观测窗口总加权飞行公里数
否	4.025545571	42.37253886	179.0645058	7176.806996	7131.13836
是	4.025226465	41.74878837	175.0708634	7066.5129	7005.866803

从中我们可以得知：

- 不同类别客户的指标平均值都不相近，易识别。
- 单从各个类别客户来讲，年龄、观察窗口内最大乘机间隔、总累计积分、观测窗口总加权飞行公里数在流失客户与非流失客户中有一定的差异，这样说明了我们的模型的准确性、有效性。

（每个客户的具体情况详见“附件 4 各客户所属类别及流失倾向.xls”）

① 发现机会：

前文已经建立客户流失预测模型，然后对在网客户进行流失倾向的评分，按倾向高低判别。并对全体客户的分群来识别出真正的挽留机会，并非流失倾向越高就越值得挽留。比如可以按照客户价值进行分群，优先考虑对中高价值客户的挽留；同时根据客户行为分群，判别出哪些客户可能已经用了竞争对手的服务，或者属于欺诈类型的客户，对这批客户的挽留可能是没有成效的，不应视为挽留机会。

② 制订策略：

经过第一个步骤，我们可以从预测名单中圈定了值得挽留的客户。但是一般来说，这批客户依然数目较大，难以逐个分析，决定采取何种挽留策略。（这也是一些厂商宣传的一对一营销，想法虽好，可是未必可行）可以对根据某个标准圈定的挽留客户进一步进行分群，将他们划分为几种类型，当然此时最好在分群模型中放入行为、人口统计学、地域等属性，然后基于这几群客户逐群制订有针对的挽留策略。（所给数据不全，此处仅作讨论）

③ 实施行动：

针对不同客户类别，应采取不同的策略来保持并提高航空上座。

- 针对重要保持客户，进行客户保持。按照总得分的排列情况，航空公司应该优先将资源投放到这类总得分较高的客户身上。
- 针对重要发展客户，进行客户发展。加强与这类客户的交流，使他们对航空公司的

会员服务、企业文化有更多的了解，提供各类会员资讯、促销信息等。

- 针对重要挽留客户，进行流失预警。我们可以定期观测重要客户的相关指标的变化情况，推测客户消费的异动状况，根据客户流失的可能性，列出客户名单，重点拜访或联系，以最有效的方式防范重要客户流失。
- 针对忠诚型一般客户，进行交叉销售。查看重要客户在非航空类合作伙伴处的里程积累情况，找出他们习惯的里程积累的方式（是否经常在合作伙伴处消费、更喜欢消费哪些类型合作伙伴的产品），对他们进行相应促销。

④ 评估效果：

在客户流失预测专题分析的试运行阶段，由于对模型预测的效果、挽留机会的识别是否准确、挽留策略的制订是否合适等方面尚未得到确认，常常会将预测名单中圈定的客户划分为两组——实施组和对照组。对前者展开挽留，对后者不采取任何行动，根据两组的流失情况来评估模型的预测效果和挽留效果。当专题分析已经基本稳定后，对照组会被取消，那么对挽留效果的评估主要依赖于客户的反馈、客户随后是否在网以及其用量的变化等来评估。

4. 参考文献

[1] 王巡，刘宇晟.航空公司消费者流失行为分析.特区经济，2012:294-296

[2] 周生宝，郭俊芳.客户流失预测模型设计与实现.计算机系统应用，2009，(5)：170-172

[3] 刘星毅，曾春华，江南雨，陈振华，韦小玲.缺失数据的处理和挑战.钦州学院学报，2008,23（6）：25—29

[4] 罗亮生，张文欣.基于常旅客数据库的航空公司客户细分方法研究.现代商业：54-55

[5] 宋晴，高彩霞，张艳茹，张源伟，王金华.Logistic 回归及倾向评分法在构建脑卒中发病概率模型中的应用.中国美容医学，2012，21（7）：123-124

[6] 张会荣，陈云.基于 SMC 模型的航空公司常旅客活跃度分析.生产力研究，2011,9:93-95

[7] 李智文，任爱国.倾向评分分层和回归分析.中国生育健康杂志，2010,21(3)：186-189