上海财经大学

SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

《数据分析与可视化》期末报告

题目：基于机器学习的电信客户流失预测

姓名：谢嘉薪

学号：2020111142

班级：20 数据科学

目录

摘要

随着客户流失问题关注度日益攀升，如何预测客户流失成为热点问题。基于此背景，本文选取 Kaggle 上的公开电信客户流失数据集的 7043 条电信客户数据及 20 个相关变量，通过建立 Logistic 模型并引入交互项和评分卡模型深入探究影响客户流失的主要特征，建立随机森林模型和 XGBoost 模型并利用 R 包自带函数输出特征重要性，发现各变量中合同期限和客户留存月份对流失情况影响最大；此外，基于上述研究，本文对各分类模型给出应用结论，并提出本次研究的不足之处，如尝试集成方法、处理类别不均等。

关键词：客户流失 逻辑回归 评分卡 随机森林 XGBoost

一、引言

（一）研究背景与研究问题

当今社会市场竞争激烈，同行业间同质化严重，客户可选择性越来越多，从而使得新增客户成本增加。研究表明发展一个新顾客成本远远超过维护老客户的成本，因此分析客户流失是各行业不可避免的话题。本文聚焦于电信客户流失预测，以帮助电信公司查寻客户流失原因，以及对即将流失的客户采取相应补救措施，从而降低客户流失率，大大提高企业利润。

电信客户流失情况有很多影响因素，如消费者个体因素、注册服务类型、账户情况等，探究客户流失的主要因素，需要建立模型与实证研究。余路（2016）运用逻辑回归、决策树、神经网络及组合模型进行客户流失预测；冀慧杰等（2021）利用 XGBoost 探索客户流失的主要因素等，均为本文进行电信客户流失预测提供了关键思路。

在此背景下，本文选取 Kaggle 上的电信客户流失数据，通过建立 Logistic 模型、随机森林模型、XGBoost 模型，进行客户流失预测。其中 Logistic 模型中引入评分卡模型，通过 IV 值探究各特征重要性并建立评分卡用于预测，随机森林和 XGBoost 模型则利用 R 包直接输出特征重要性排名并进行可视化。根据机器学习的方法，基于电信客户大数据，运营商便能预测客户流失的可能性大小，及时获取“待流失”的客户名单，可尽早通过提供优惠等途径提高用户粘性，挽回用户，降低客户流失率。

（二）数据获取与变量说明

本文数据集来自 Kaggle 平台的公开数据——Telco Customer Churn[1]，它包括 7043 条电信客户的数据、20 个特征、1 个分类标签，其中特征有一列为样本编号，无意义，予以剔除，剩余 16 列特征为分类型变量，3 列为数值型变量；分类标签“客户流失（Churn）”为本文的目标变量，分为“流失客户”和“现有客户（非流失）”两类，代表上个月的客户流失情况。

该数据集中含 11 个缺失值，相对样本可忽略不计，因此本文选择对缺失值所在的样本进行剔除。此外，客户注册服务类型变量分电话服务和互联网服务两方面，其中细分变量如“是否有多条线路”建立在有电话服务之上，“在线安全服务”、“在线备份”、“设备保护”、“技术支持”、“媒体电视服务”、“媒体电影服务”建立在互联网服务之上，以上变量均有三个水平：“是”、“否”、“无电话/互联网服务”，其中“无电话/互联网服务”直接导致无对应细分服务内容，为便于后续建模和分析，将其与“否”合并为同一水平。综上，变量说明如表 1 所示。

表 1 变量说明

指标类型	变量名称	变量含义	取值情况
自变量 客户个人信息	Gender	性别	男，女
	Senior Citizen	是否是老人	是，否
	Partner	是否有朋友	是，否

[1] 数据来源：https://www.kaggle.com/datasets/blastchar/telco-customer-churn

	Dependents	是否有家属	是，否
自变量 客户注册服务	Phone Service	是否有电话服务	是，否
	Multiple Lines	是否有多条线路	是，否
	Internet Service	互联网服务类型	无，拨号上网，光纤
	Online Security	是否有在线安全服务	是，否
	Online Backup	是否有在线备份服务	是，否
	Device Protection	是否有设备保护服务	是，否
	Tech Support	是否有技术支持服务	是，否
	Streaming TV	是否有媒体电视服务	是，否
	Streaming Movies	是否有媒体电影服务	是，否
自变量 客户账户信息	Tenure	客户留存月数	1-72
	Contract	合同期限类型	逐月，一年期，两年期
	Paperless Billing	是否有纸质账单	是，否
	Payment Method	付款方式	银行转账（自动） 信用卡转账（自动） 电子支票 纸质支票
	Monthly Charges	每月收取金额	18.25-118.75
	Total Charges	向客户收取的总金额	18.8-8684.8
因变量	Churn	是否流失	是，否

二、探索性数据分析

（一）自变量与因变量的关系

本文研究的自变量包括 16 个分类变量和 3 个连续变量，为判断自变量对因变量是否有影响，绘制分类变量的堆积图并对其依次做卡方检验，若拒绝原假设，表明自变量与因变量有显著的相关性，即自变量会对因变量有影响。连续变量则绘制箱线图并进行单因素方差分析。

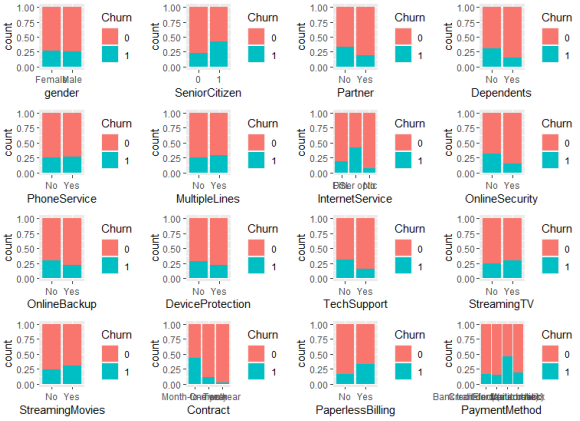


图 1 分类变量与因变量的条形图

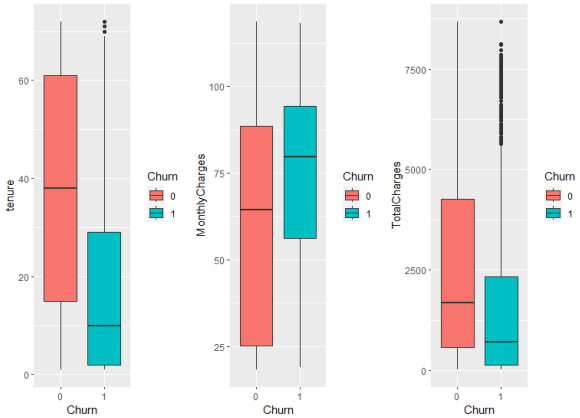


图 2 连续变量与因变量的箱线图

对于分类变量，观察图 1 看出，不同性别的流失客户占比、是否有电话服务的流失客户占比非常相近，即这两个因素几乎不会对客户流失产生影响，推测是因为电信服务属于基础服务，不存在审美等存在男女差异的属性，而电话服务附加价值较低，且在互联网时代逐渐可有可无。相反，合同期限类型为两年付的客户流失占比远低于月付的，是因为年费客户一方面出于违约成本不更换运营商，形成绑定关系，另一方面越忠诚的客户越倾向于订购长期合同，符合基础认知；无网络服务的用户流失概率远低于有光纤服务的用户，这是因为对于有网络需求的客户来说，不同运营商的网络服务质量和优惠力度等众多因素都会造成流动，而不需要网络的客户则相对流动性较弱。

对于连续变量，观察图 2 可以看出，在因变量的两个水平上，连续变量的平均水平差异均较大，即都会对客户是否流失产生一定影响，其中流失客户的留存月份平均水平较低，考虑到观测流失情况

的时点均为上个月，初步判断新用户更容易流失，可能是行业同质化所致。

为使后续模型建立和分析更有效，对卡方检验和单因素方差分析中 P-value>0.001 的变量予以剔除。各变量的 P-value 如图 3 所示，基于此剔除性别、是否有电话服务两个变量，其次是否有多条线路是基于电话服务的具体内容，若不研究电话服务则该变量也无研究意义，予以剔除。

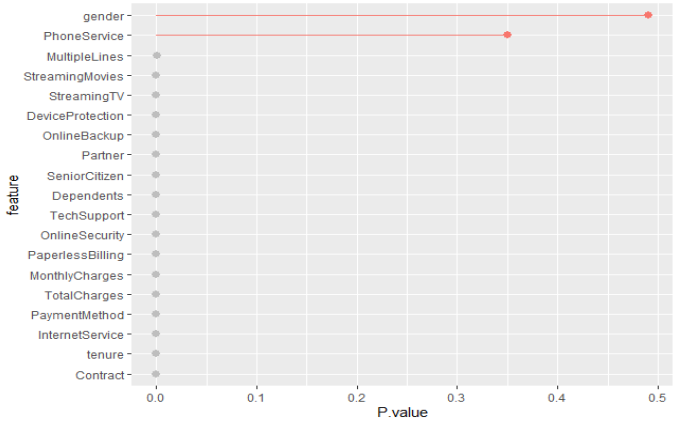


图 3 分类变量与因变量的卡方检验/方差分析

（二）自变量之间的关系

对于分类变量，上一小节中推测：有网络需求，尤其是光纤的客户流动性较强，不需要网络的客户流动性较弱。为验证这一推测，画出合同期限类型在互联网服务类型上的堆积图如图 4，可以发现：光纤客户签订月付合同占比远大于其他两种合同期限，无网络服务的客户则更倾向于签订年付合同。

对于连续变量，如图 5 所示，将 3 个连续变量绘制成散点图，分类标签映射为颜色，进而可以看出变量的交互作用对因变量是否有影响，具体将在 Logistic 模型建立时展开说明；同时通过对角线上不同响应水平的频数分布存在差异可以印证连续变量与因变量不独立，具有研究意义。

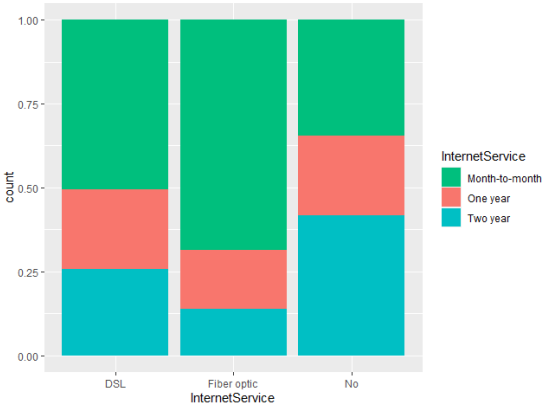


图 4 互联网服务与合同期限的关系

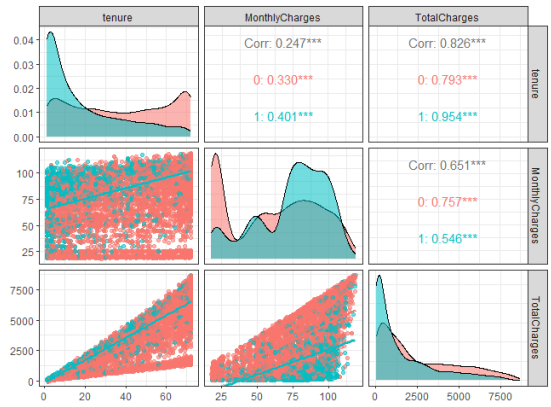


图 5 连续变量之间的关系

三、研究思路与方法简介

本文将数据集按 7:3 划分为训练集和测试集 [2]，对于需要迭代求最优参数的模型，将训练集再划分为新训练集和验证集，然后通过 Logistic 回归模型、随机森林模型、XGBoost 模型探究影响客户流失的重要特征并对客户流失情况进行预测，最后通过 ROC 曲线、AUC 等评价指标评估模型效果，其中篇幅原因混淆矩阵在附录给出，文中直接展示准确率、召回率和精确率。

（1）Logistic 回归模型

假设有数据集$\{(y_i, x_i)\}_{i=1}^n$，其中$y_i \in \{0,1\}$，Logistic 回归尝试估计$p(y_i = 1|x_i) \triangleq p(x_i)$。考虑线性组合$w^T x_i + b$，一个实际的问题是：$p(x_i)$的取值范围为(0,1)，而$w^T x_i + b$可能取到一切实数，即它们的值域不同，考虑 Logit 变换：

[2] 本文所有具有随机性的步骤均设定种子为 2020111142

$$logit\big(p(x_i)\big) = \log\frac{p(x_i)}{1-p(x_i)}, i = 1, \dots, n$$

直接计算可得：

$$p(x_i) = \frac{1}{1+e^{-x_i^T\beta}}$$

通常将$p(x_i) > 0.5$的实例划分为正例，反之为反例，但本文以准确率为评价指标，通过迭代寻求最优分类阈值，进而提高模型精度。

（2）评分卡模型

信用评分卡在信用风险评估与控制领域有广泛的使用，其原理对连续变量离散化后和分类变量一起进行 WOE 编码，再运用 logistic 回归模型进行二分类的广义线性模型。对于本问题，采用 EM 算法对连续变量进行分箱。

EM（期望最大化）聚类并不计算距离，而是计算概率，用一个给定的多元高斯概率分布模型来估计出一个数据点属于一个聚类的概率，主要由两步交替进行：

E-step：对于每一个数据点，我们要计算其属于其中每个聚类的概率作为权重，对于那种可能会出现一个点属于 2 个或多个聚类的情况，就需要建立一个对聚类的概率分布。

M-step：这一步骤主要是利用上一步计算的权重来估计每个聚类的有关参数：每一个数据点以概率作权重，计算每一个聚类的均值和方差，进而求取聚类的总体概率或极大似然。

分箱完成后即可进行 WOE（Weight of Evidence）编码。记因变量为$y_i \in \{0,1\}$，y = 0 的频数为$N0$，y = 1 的频数为$N1$，对于一个分类变量X_i,设它有n_i个水平，在第 K 个水平上$y=0$ 的频数为$N_{K_0}^{X_i}$，$y=1$ 的频数为$N_{K_1}^{X_i}$，可定义该变量在该水平上的 WOE 值，并算出对应 IV 值：

$$WOE(X_i) = ln\left(\frac{N_{K_1}^{X_i}/N_{K_0}^{X_i}}{N_1/N_0}\right)$$

$$IV = \sum_i^n \left(\frac{N_{K_1}^{X_i}}{N_{K_0}^{X_i}} - \frac{N_1}{N_0}\right) WOE(X_i)$$

WOE 值刻画了该变量的当前水平对因变量起到的影响的方向和大小，WOE 值大于 0 表明当前水平$y=1$ 的比例超过整体比例，而 WOE 值越大，表明在当前水平中$y=1$ 的比例越大。IV 值则是一种衡量自变量预测能力的指标，我们可以根据 IV 值对变量进行排序，观察每个变量的预测能力大小。

（3）随机森林模型

随机森林是将多个决策树组合而成，在训练决策树时需要衡量拆分选取的好坏、停止拆分的准则、每个叶节点所预测的类。从根节点开始，遍历每一个变量的每一种拆分方式，在不纯度准则下找到最好拆分；将根节点分拆出两个子节点，并对子节点再次运用上步；重复上述步骤，直到达到事先给定的停止条件。随机森林的预测结果是由内部所有决策树的预测结果投票而得。随机森林模型还能计算特征重要性，表示特征对预测结果的影响程度。

（4）XGBoost 模型

XGBoost 是一种提升树算法，即构建多个决策树，当构建下一颗树的时候需要对上一颗树进行评估，按照表现不佳的部分训练下一颗树，依次类推。XGBoost 在 GBDT 的基础上增加了正则化步骤，速度比 GBDT 更快、精度更高，且基本不会过拟合，常被称作基准方法。同时 XGBoost 也能输出特征的重要性，其评价指标包括信息增益、特征所涉及的样本相对数量、特征在生成树中所用的次数。

四、实证分析

（一）Logistic 回归模型

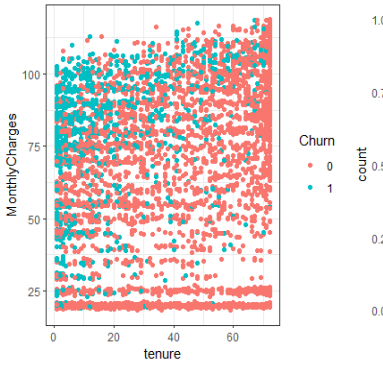
模型建立分两部分，首先对交互变量探索并代入基本 Logistic 模型，输出回归系数初步探索特征

重要性；其次是运用评分卡模型进行改进形成最终模型，输出 IV 值并评估预测效果。

1、交互项变量探索

由于在探索性数据分析中发现，连续变量之间可能存在交互作用，本部分通过绘制客户留存月数、向客户收取的总金额分别和每月收取金额的散点图，并将是否流失映射到不同颜色，通过观察找出可能存在交互作用的区域，再将可能存在交互作用的变量放入模型中拟合，查看系数是否显著。

对于客户留存月数和每月收取金额的交互作用，通过图 6 散点图可以看出：在留存月数少于 15，且月缴费大于 75 的区域内，观测结果集中为流失，进一步可观察堆积图，发现是否流失在区域内外差别较大，故考虑新的交互项 tenure<=15&MonthlyCharges>=75。此处引入交互项的方法采用引入一个哑变量：{ tenure<=15&MonthlyCharges>=75}==True，放入模型后系数显著。



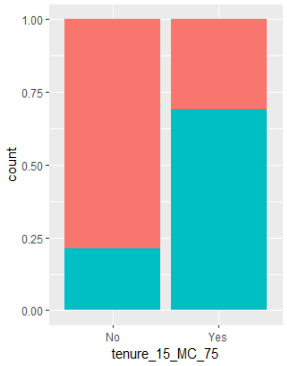
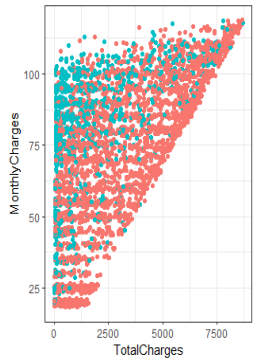


图 6 客户留存月数和每月收取金额的交互作用



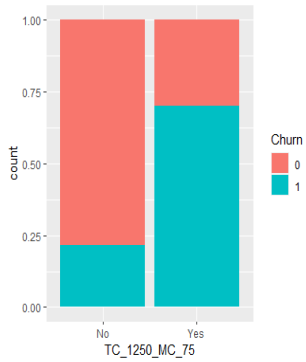


图 7 向客户收取的总金额和每月收取金额的交互作用

对于向客户收取的总金额和每月收取金额的交互作用，通过图 7 散点图可以看出：在总金额小于 1250，月缴费大于 75 的区域内，观测结果集中为流失，进一步观察堆积图发现是否流失在区域内外差别较大，故考虑新交互项 TotalCharges<=1250&MonthlyCharges>=75。同样引入一个哑变量：{TotalCharges<=1250&MonthlyCharges>=75}==True，放入模型后系数显著。

2、基本模型建立

利用 glm 函数，对训练集的目标变量 Churn 及所有自变量和交互项建立 Logistic 回归模型，输出回归系数并绘制条形图，能清晰看出各变量对目标变量的影响程度，如图 8 所示。可以发现，互联网服务类型和合同期限是对于客户是否流失最显著的特征，其中有光纤服务的客户更容易流失，无互联网服务和两年期合同的客户更不易流失，再次验证初步猜想。

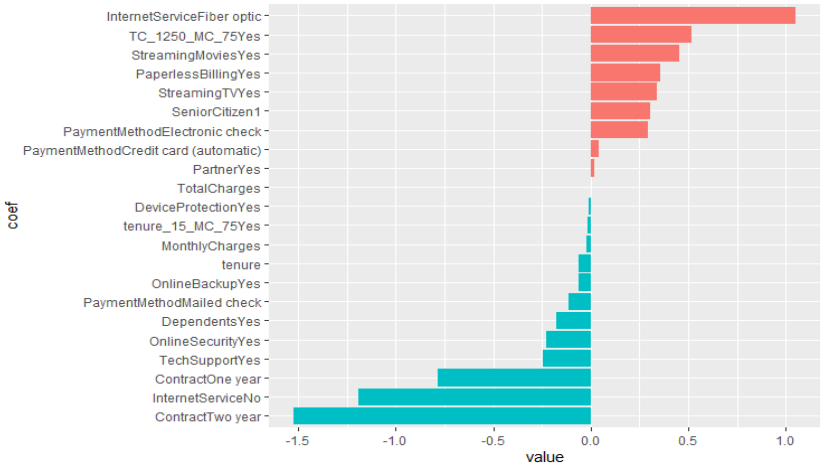


图 8 初步拟合回归系数

3、评分卡模型

首先利用 Mclust 函数对连续变量进行分箱，选择簇数为 4，部分聚类结果如图 9 所示。用样本点所在的类别作为样本点的值，将这三个连续变量变为分类变量，依次与因变量做卡方检验，结果显示全部拒绝原假设（P<0.001），即分箱后的变量与因变量不独立，效果理想。其次是对目前剩余的 18 个分类变量（连续变量已分箱）进行 WOE 编码，用 WOE 值作为自变量相应水平的替代值。基于此绘制相关性矩阵，以初步筛选变量，简洁起见，图 10 中仅展示相关性高的变量。

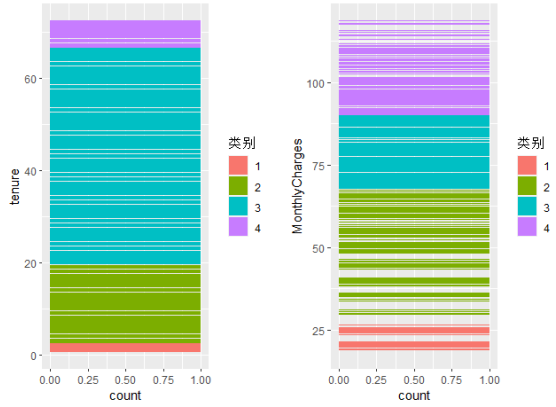


图 9 聚类结果可视化

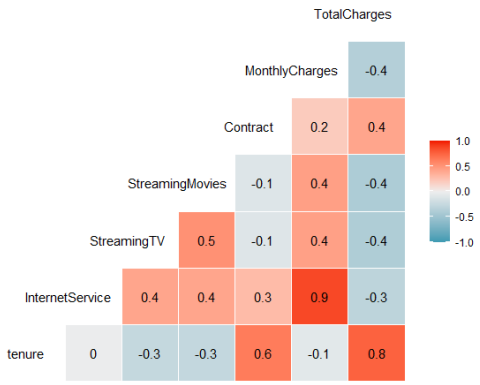


图 10 相关性矩阵（相关性较高变量）

可以发现，是否有互联网服务与月缴费高度相关，不难推测是因为网络费用较高导致，符合实际，因此可删去其一，此处保留前者；客户留存时间与总缴费高度相关，但考虑到存在长期留存但服务较少而总缴费较少的特殊情况，本文忽略此共线性，保留二者。对剩余的 17 个分类变量，采用训练集建立 Logistic 回归模型，回归系数见表 2，其中较为反常的是，WOE 编码后，WOE 的值越大，$P(y=1)$越大，因此回归系数应该都是非负数，但其中是否有朋友、留存月数和月缴费的交互项系数均小于 0，且都很小，推测是自变量列的复共线性导致的，因此将它们删除，再拟合 Logistic 回归模型，调整后的回归系数见附录 1。同时将 IV 值排序并绘制条形图，如图 11 所示。

表 2 回归系数表

变量	回归系数	变量	回归系数
截距项	-1.05158	TechSupport	0.23475
SeniorCitizen	0.38970	StreamingTV	1.26618
Partner	-0.05255	StreamingMovies	1.60262
Dependents	0.17018	Contract	0.50606
tenure	0.38476	PaperlessBilling	0.41550
InternetService	0.84463	PaymentMethod	0.21424
OnlineSecurity	0.22925	TotalCharges	0.80524
OnlineBackup	0.20982	tenure_15_MC_75	-0.07289
DeviceProtection	0.01224	TC_1250_MC_75	0.23158

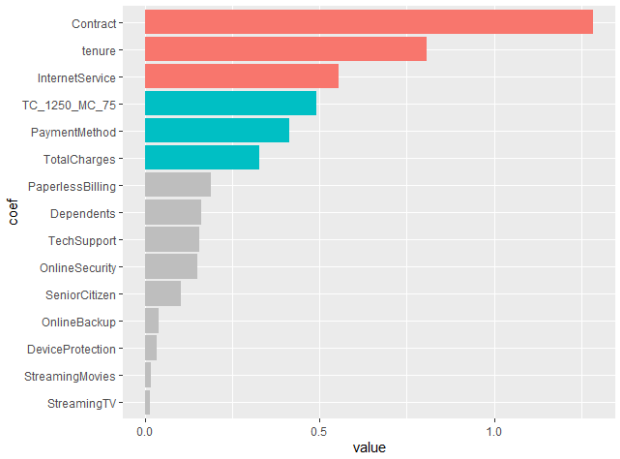
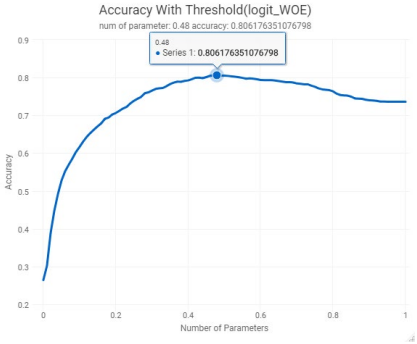


图 11 IV值排序

模型结果可从两方面应用，一是基于评分卡思想，将自变量的 WOE 值与回归系数相乘再扩大 100 倍作为评分卡，见附录 2，将电信客户特征代入表中并把评分相加，总分越高表面其流失的概率越大；二是观察IV值，IV值介于 0.3~0.49 之间说明具有高预测能力，≥0.5 说明具有极高的预测能力。根据上图结果，合同期限、留存月份、是否有互联网服务具有极高的预测能力，总缴费及其与月缴费的交互项、支付方式具有高预测能力。因此这 6 个指标可以作为预测客户流失的重要因素，需重点关注。

4、模型评估

模型拟合后需要确定最佳阈值，为此通过迭代确定最优阈值为 0.48，如图 12 所示，使用最优分类阈值对测试集进行预测得到的 ROC 曲线如图 13 所示，此时 AUC 为 0.846，准确率为 0.808，召回率为 0.874，精确率为 0.865，拟合效果良好。



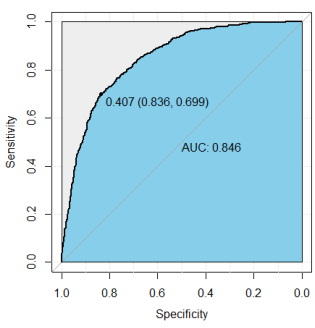
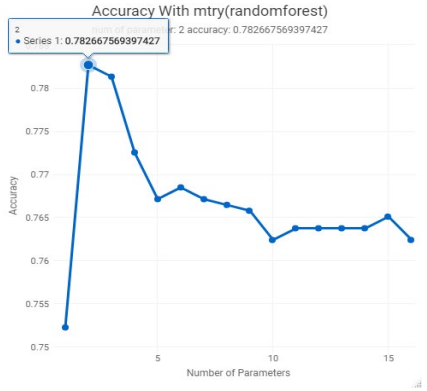


图 12 准确率随阈值变化情况(Logistic)　　图 13 Logistic 预测 ROC 曲线

（二）随机森林模型

通过 randomForest 函数建立随机森林模型。首先寻找最优参数 mtry，即指定节点中用于二叉树的最佳变量个数，通过循环分别构造 mtry 为 1-16 时的随机森林，通过图 14 选取在验证集上预测准确率最高的 mtry=2；其次对于 ntree 的选取，由于不建议过小，设定为 1000，根据图 15 发现随着树个数的增加误差逐渐减小，在 1000 处几乎无变化，故 ntree=1000 合理。



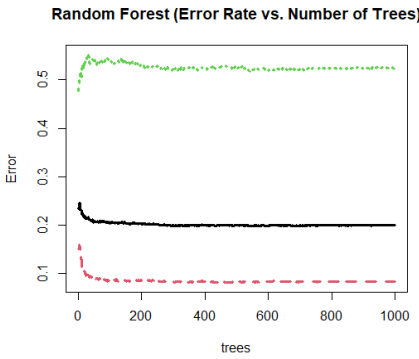
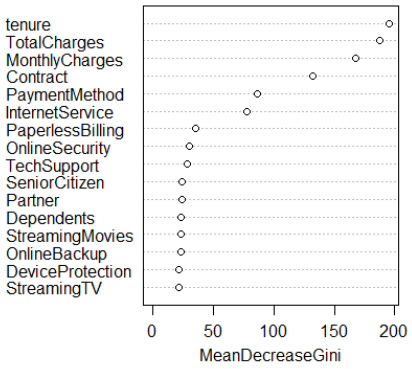


图 14 准确率随 mtry 变化情况　　图 15 误差随 ntree 变化情况

通过 varImpPlot 函数，得到如图 15 所示通过拆分特定特征的情况下节点非纯度平均减少量度量的特征重要性排序，可以看出客户留存月份、缴费情况、合同期限重要性均超过 100，说明其对客户流失的影响程度较大。模型建立完毕后对测试集进行预测，得到的 ROC 曲线如图 17 所示，此时 AUC 为 0.841，准确率为 0.804，召回率为 0.914，精确率为 0.834，拟合效果良好。



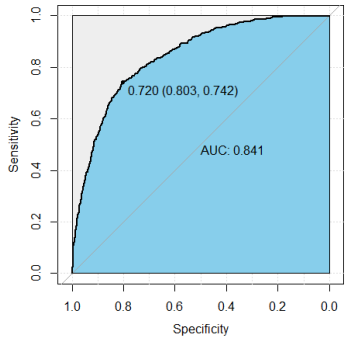


图 16 特征重要性排序（RF）　　图 17 随机森林预测 ROC 曲线

（三）XGBoost 模型

首先通过 Matrix 包和 xgb.Dmatrix 函数构造模型所需要的数据对象，然后用 xgboost 函数建立模型，设置 max_depth 为 2，学习率为 0.5，在训练集上进行训练，并通过 importance 函数输出特征重要性，此处绘制基于信息增益的重要性排序，如图 18 所示，最重要的两个特征是合同期限和客户留存月份，与 WOE 编码结果相同。

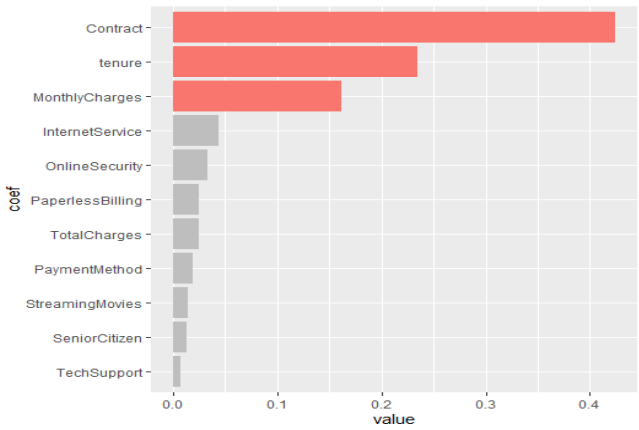


图 18 特征重要性排序（XGBoost）

由于模型输出的分类结果同 Logistic 模型也是概率值，因此同样可通过迭代寻求最优分类阈值，确定为 0.49，如图 19 所示，使用最优分类阈值对测试集进行预测得到的 ROC 曲线如图 20 所示，此时 AUC 为 0.849，准确率为 0.806，召回率为 0.912，精确率为 0.836，拟合效果良好。

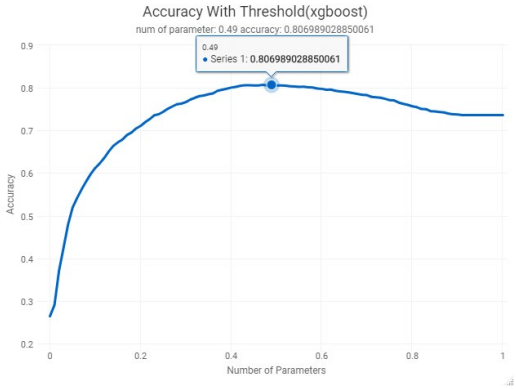


图 19 准确率随阈值变化情况(XGBoost)

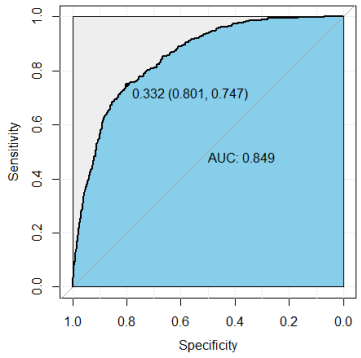


图 20 XGBoost 预测 ROC 曲线

五、结论

（一）模型比较

除文中所用模型外，还建立了 SVM 模型，但考虑到效果一般且未输出特征重要性，仅在此列出。

表 3 模型评价指标表

模型	AUC	准确率	召回率	精确率
Logistic	0.846	**0.808**	0.874	**0.865**
随机森林	0.841	0.804	0.914	0.834
XGBoost	**0.849**	0.806	0.912	0.836
SVM	0.683	0.800	**0.916**	0.829

（1）由于研究问题是二分类问题，AUC 往往作为关键评价指标，从上表中可知前三种分类器效果接近，其中 XGBoost 效果最佳，但解释性最差，而 Logistic 具有最强的解释性，其在准确性和精确性上表演最好，可作为最优分类器。

（2）在特征重要性探索方面，Logistic 回归引入了评分卡模型，具有最大的应用价值，由 Logistic 回归方程和 WOE 的映射规则，不难算出：总分每增加 1 分，客户流失的对数“优势”增加 1%；当总分超过 429 分时，客户流失的预测概率达到 90%以上。随机森林和 XGBoost 也给出了特征重要性，结论可与评分卡模型相呼应，同时三者均可验证探索性数据分析时的猜想。

（二）研究结论

基于上述对电信客户流失预测模型的研究，发现合同期限、客户留存月份对客户流失影响最大，这提醒运营商可通过对长期合同采取较大优惠力度与客户形成绑定关系，重点关注新用户。其次互联网服务类型也较为重要，对于光纤入户的新客户应给予重点关注。最后，通过评分卡和客户大数据可以预测出客户流失的概率（分数），对于此类客户应及时回馈，采取优惠策略留住该类人群。

（三）不足之处

（1）虽然本文研究了多种分类模型的效果，但采用的是单一模型，可以尝试借助 Stacking 集成方法融合几种模型，探究组合模型的效果是否有显著提升。

（2）由于流失客户占少数，导致存在样本类别不均的情况，对于 Logistic 模型影响较小，但对随机森林和 XGBoost 的效果有一定限制，可考虑用 SMOTE 采样法和随机抽样法进行改进。

参考文献

[1]余路.电信客户流失的组合预测模型[J].华侨大学学报(自然科学版),2016,37(05):637-640.
[2]夏国恩,金炜东.基于支持向量机的客户流失预测模型[J].系统工程理论与实践,2008(01):71-77.
[3]于小兵,曹杰,巩在武.客户流失问题研究综述[J].计算机集成制造系统,2012,18(10):2253-2263.DOI:10.13196/j.cims.2012.10.125.yuxb.017.
[4]陈可.基于B-SMOTE1-XGBoost预测电信客户流失[J].郑州师范教育,2022,11(04):21-26.
[5]冀慧杰,倪枫,刘姜,陆祺灵,张旭阳,阙中力.基于XGB-BFS特征选择算法的电信客户流失预测[J].计算机技术与发展,2021,31(05):21-25.
[6]王雷,陈松林,顾学道.客户流失预警模型及其在电信企业的应用[J].电信科学,2006(09):47-51.

附录

附录 1 Logistic 回归系数表

变量	回归系数	变量	回归系数
截距项	-1.05116	TechSupport	0.23511
SeniorCitizen	0.39362	StreamingTV	1.25545
Dependents	0.15228	StreamingMovies	1.59174
tenure	0.37368	Contract	0.50496
InternetService	0.84397	PaperlessBilling	0.41562
OnlineSecurity	0.23073	PaymentMethod	0.21415
OnlineBackup	0.20847	TotalCharges	0.80882
DeviceProtection	0.01195	TC_1250_MC_75	0.16301

附录 2 评分卡

变量	取值水平	评分
是否是老人	是	26.34
	否	-6.00
是否有家属	是	-10.84
	否	3.51
互联网服务类型	拨号上网	-35.31
	光纤	56.54
	无服务	-116.69
是否有在线安全服务	是	-15.52
	否	5.21
是否有在线备份服务	是	-5.82
	否	2.79
是否有设备保护服务	是	-0.31
	否	0.15
是否有技术支持服务	是	-16.07
	否	5.40
是否有媒体电视服务	是	18.58
	否	-12.10
是否有媒体电影服务	是	25.31
	否	-16.94
客户留存月数	第一类型	52.22
	第二类型	20.15
	第三类型	-20.18
	第四类型	-75.07
合同期限类型	逐月	37.14
	一年期	-56.43
	两年期	-129.26

是否有纸质账单	是	13.35
	否	-24.83
付款方式	银行转账（自动）	-12.50
	信用卡转账（自动）	-13.12
	电子支票	17.16
	纸质支票	-8.40
向客户收取的总金额	第一类型	92.92
	第二类型	22.74
	第三类型	-13.93
	第四类型	-45.26
总缴费<1250 & 月缴费>75	是	30.50
	否	-4.45

附录 3 混淆矩阵

Logistic 模型

混淆矩阵		真实值	
		0	1
预测值	0	1346	210
	1	195	359

随机森林模型

混淆矩阵		真实值	
		0	1
预测值	0	1408	280
	1	133	289

XGBoost 模型

混淆矩阵		真实值	
		0	1
预测值	0	1406	275
	1	135	294

附录 4 代码

```
library(tidyverse)
library(GGally)
library(patchwork)
library(mclust)
library(caret)
library(pROC)
library(highcharter)
library(e1071)
library(randomForest)
library(smotefamily)
library(xgboost)
library(Matrix)
rm(list = ls())

data <- read_csv("WA_Fn-UseC_-Telco-Customer-Churn.csv")
summary(data)
data <- data[,-1] #删除 ID 列
table(data$Churn) #样本不均
data$Churn <- factor(data$Churn,labels = c("0","1"))
#响应变量因子化
str(data)
```

```
sum(is.na(data))
data <- na.omit(data)

### 个别值处理
data$MultipleLines[data$MultipleLines == 'No phone service'] <- 'No'
data$OnlineSecurity[data$OnlineSecurity == 'No internet service'] <- 'No'
data$OnlineBackup[data$OnlineBackup == 'No internet service'] <- 'No'
data$DeviceProtection[data$DeviceProtection == 'No internet service'] <- 'No'
data$TechSupport[data$TechSupport == 'No internet service'] <- 'No'
data$StreamingTV[data$StreamingTV == 'No internet service'] <- 'No'
data$StreamingMovies[data$StreamingMovies == 'No internet service'] <- 'No'

### 数据类型处理
data <- data %>%
  mutate(SeniorCitizen = as.character(SeniorCitizen)) %>%
  mutate_if(is.character,factor,ordered=F)

sapply(data,class)

### 卡方检验 分类变量
# gender  不显著
pv1 <- chisq.test(table(data$gender,data$Churn))$p.value
p1 <- ggplot(data = data) +
  geom_bar(mapping = aes(x = gender, fill = Churn), position = "fill")
pv$p.value

# SeniorCitizen
pv2 <- chisq.test(table(data$SeniorCitizen,data$Churn))$p.value
p2 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = SeniorCitizen, fill = Churn), position = "fill")

# Partner
pv3 <- chisq.test(table(data$Partner,data$Churn))$p.value
p3 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = Partner, fill = Churn), position = "fill")

# Dependents
pv4 <- chisq.test(table(data$Dependents,data$Churn))$p.value
p4 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = Dependents, fill = Churn), position = "fill")

# PhoneService  不显著
pv5<- chisq.test(table(data$PhoneService,data$Churn))$p.value
p5 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = PhoneService, fill = Churn), position = "fill")

# MultipleLines  无意义
pv6<- chisq.test(table(data$MultipleLines,data$Churn))$p.value
p6 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = MultipleLines, fill = Churn), position = "fill")

# InternetService
pv7<- chisq.test(table(data$InternetService,data$Churn))$p.value
p7 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = InternetService, fill = Churn), position = "fill")

# OnlineSecurity
pv8<- chisq.test(table(data$OnlineSecurity,data$Churn))$p.value
p8 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = OnlineSecurity, fill = Churn), position = "fill")

# OnlineBackup
pv9<- chisq.test(table(data$OnlineBackup,data$Churn))$p.
```

```
value
p9 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = OnlineBackup, fill = Churn), position = "fill")

# DeviceProtection
pv10<-chisq.test(table(data$DeviceProtection,data$Churn))$p.value
p10 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = DeviceProtection, fill = Churn), position = "fill")

# TechSupport
pv11<-chisq.test(table(data$TechSupport,data$Churn))$p.value
p11 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = TechSupport, fill = Churn), position = "fill")

# StreamingTV
pv12<-chisq.test(table(data$StreamingTV,data$Churn))$p.value
p12 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = StreamingTV, fill = Churn), position = "fill")

# StreamingMovies
pv13<-chisq.test(table(data$StreamingMovies,data$Churn))$p.value
p13 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = StreamingMovies, fill = Churn), position = "fill")

# Contract
pv14<-chisq.test(table(data$Contract,data$Churn))$p.value
p14 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = Contract, fill = Churn), position = "fill")

# PaperlessBilling
pv15<-chisq.test(table(data$PaperlessBilling,data$Churn))$p.value
p15 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = PaperlessBilling, fill = Churn), position = "fill")

# PaymentMethod
pv16<-chisq.test(table(data$PaymentMethod,data$Churn))$p.value
p16 <- ggplot(data=data) +
  geom_bar(mapping = aes(x = PaymentMethod, fill = Churn), position = "fill")

p1+p2+p3+p4+p5+p6+p7+p8+p9+p10+p11+p12+p13+p14+p15+p16

### 单因素方差分析
# tenure
tenure.aov <- aov(data$tenure~data$Churn)
pv17<-summary(tenure.aov)[[1]][["Pr(>F)"]][1]

# MonthlyCharges
MonthlyCharges.aov <- aov(data$MonthlyCharges~data$Churn)
pv18<-summary(MonthlyCharges.aov)[[1]][["Pr(>F)"]][1]

# TotalCharges
TotalCharges.aov <- aov(data$TotalCharges~data$Churn)
pv19<-summary(TotalCharges.aov)[[1]][["Pr(>F)"]][1]

p17 <- ggplot(data=data) +
  geom_boxplot(mapping = aes(x = Churn, y = tenure, fill=Churn))
p18 <- ggplot(data=data) +
  geom_boxplot(mapping = aes(x = Churn, y = MonthlyCharges, fill=Churn))
p19 <- ggplot(data=data) +
  geom_boxplot(mapping = aes(x = Churn, y = TotalCharges, fill=Churn))

p17+p18+p19

# p-value plot
pv<-
```

```
c(pv1,pv2,pv3,pv4,pv17,pv5,pv6,pv7,pv8,pv9,pv10,
pv11,pv12,pv13,pv14,pv15,pv16,pv18,pv19)
p.frame <- data.frame(feature = names(data[-
20]),P.value = pv)
p.frame <- p.frame[order(p.frame$P.value),]
p.frame$feature <- factor(p.frame$feature,levels =
p.frame$feature)
ggplot(data = p.frame,aes(x = feature, y = P.value)) +
  geom_pointrange(aes(ymin=0,ymax=P.value),
                  size = 0.5,
                  color =
ifelse(p.frame$P.value>0.001,'#F8766D','grey'))+
  coord_flip()

### 剔除不显著
data <- data %>%
  dplyr::select(-gender,-PhoneService,-
MultipleLines)

data2 <- data

ggplot(data=data) +
  geom_bar(mapping = aes(x = Contract, fill =
InternetService), position = "fill")+
  scale_fill_manual('InternetService',values =
c('Fiber
optic'='#F8766D','DSL'='#00BF7D','No'='#00BFC4'))

ggplot(data=data) +
  geom_bar(mapping = aes(x = InternetService, fill
= Contract), position = "fill")+
  scale_fill_manual('InternetService',values =
c('Month-to-month'='#00BF7D','One
year'='#F8766D','Two year'='#00BFC4'))

### 连续变量交互作用
dataCON <- data %>%
  dplyr::select(where(is.numeric),Churn)

ggpairs(dataCON[-
4],aes(color=dataCON$Churn,alpha=0.75),lower =
list(continuous="smooth"))+
  theme_bw()



### Logistic
### 划分数据集
set.seed(2020111142)
ind <- sample(1:nrow(data),size = 0.7*nrow(data))
train <- data[ind,]
test <- data[-ind,]

### 交互变量探索
#tenure vs MonthlyCharges
p20 <- ggplot(train,aes(x=tenure, y=MonthlyCharges,
color=Churn))+
  geom_point()+
  theme_bw()
# train 添加交互项
train$tenure_15_MC_75 <-
factor(ifelse(train$tenure<=15 &
train$MonthlyCharges>=75,'Yes','No'))
chisq.test(table(train$tenure_15_MC_75,train$Chur
n))
p21 <-ggplot(data=train) +
  geom_bar(mapping = aes(x = tenure_15_MC_75,
fill = Churn), position = "fill")
p20+p21

# test 添加交互项
test$tenure_15_MC_75 <-
factor(ifelse(test$tenure<=15 &
test$MonthlyCharges>=75,'Yes','No'))

#tenure vs MonthlyCharges
p22 <- ggplot(train,aes(x=TotalCharges,
y=MonthlyCharges, color=Churn))+
  geom_point()+
  theme_bw()
# train 添加交互项
train$TC_1250_MC_75 <-
factor(ifelse(train$TotalCharges<=1250 &
train$MonthlyCharges>=75,'Yes','No'))
chisq.test(table(train$TC_1250_MC_75,train$Churn
))
p23 <- ggplot(data=train) +
  geom_bar(mapping = aes(x = TC_1250_MC_75,
fill = Churn), position = "fill")
p22+p23

# test 添加交互项
test$TC_1250_MC_75 <-
factor(ifelse(test$TotalCharges<=1250 &
```

```
test$MonthlyCharges>=75,'Yes','No'))

# 建模
fit.logit.inter <- glm(Churn~.,train,family = binomial(link = "logit"))
summary(fit.logit.inter)

# 系数可视化
coefs <- coef(fit.logit.inter)[-1]
coef.frame <- data.frame(coef = names(coefs),value = coefs)
coef.frame <- coef.frame[order(coef.frame$value),]
coef.frame$coef <- factor(coef.frame$coef,levels = coef.frame$coef)
ggplot(data = coef.frame,aes(x = coef, y = value)) +
  geom_bar(stat = 'identity',
           fill = ifelse(coef.frame$value>0,'#F8766D','#00BFC4'), # 根据 y 值的正负设置颜色
           width = 0.9)+
  coord_flip()



# 确定最佳阈值
pred.train.logit.inter.p <- predict(fit.logit.inter, train, type = 'response')
thresholds <- seq(0,1,by=0.01)
acc.test <- numeric()
accuracy1 <- NULL
accuracy2 <- NULL
for(i in 1:length(thresholds)){
  pred.train.logit.inter <- ifelse(pred.train.logit.inter.p<thresholds[i],"0","1")
  accuracy1 <- confusionMatrix(factor(pred.train.logit.inter,levels = c("0","1")),train$Churn)
  accuracy2[i] <- accuracy1$overall[1]
}
acc.test <- data.frame(p=thresholds,cnt=accuracy2)
opt.p <- subset(acc.test,cnt==max(cnt))[1,]
sub <- paste("num of parameter:",opt.p$p," accuracy:", opt.p$cnt)
sub

hchart(acc.test,'line',hcaes(p,cnt))%>%
  hc_title(text='Accuracy With Threshold(logit_all)')%>%
  hc_subtitle(text=sub)%>%
  hc_add_theme(hc_theme_google())%>%
  hc_xAxis(title=list(text = 'Number of Parameters'))%>%
  hc_yAxis(title=list(text = 'Accuracy'))

# 使用最优阈值分类
pred.logit.inter.p <- predict(fit.logit.inter,newdata = test,type = 'response')
pred.logit.inter <- ifelse(pred.logit.inter.p < opt.p$p,'0','1')
confusionMatrix(data = factor(pred.logit.inter,levels = c('0','1')),reference = test$Churn)

roc.logit.inter <- roc(test$Churn,pred.logit.inter.p,quiet=T)
plot(roc.logit.inter,print.auc=T,auc.polygon=T,grid=c(0.1,0.2),max.auc.polygon=T,auc.polygon.col='skyblue',print.thres=T)


### 对训练集连续变量分箱
trainCON <- train %>%
  dplyr::select(where(is.numeric),Churn)

set.seed(2020111142)
tenure.clust <- Mclust(trainCON$tenure,G=4) #(v,4)
summary(tenure.clust, parameters = T)
p24 <- ggplot(cbind(train, 类别=factor(tenure.clust$classification)),aes(y=tenure,fill=类别))+
  geom_bar(position = 'fill')
train$tenure <- factor(tenure.clust$classification)
table(factor(tenure.clust$classification),train$Churn)%>%chisq.test()

MonthlyCharges.clust <- Mclust(trainCON$MonthlyCharges,G=4) #(v,4)
summary(MonthlyCharges.clust, parameters = T)
p25 <- ggplot(cbind(train, 类别=factor(MonthlyCharges.clust$classification)),aes(y=MonthlyCharges,fill=类别))+
  geom_bar(position = 'fill')
train$MonthlyCharges <- factor(MonthlyCharges.clust$classification)
table(factor(MonthlyCharges.clust$classification),tra
```

```
in$Churn)%>%chisq.test()

TotalCharges.clust <-
Mclust(trainCON$TotalCharges,G=4)  #(v,4)
summary(TotalCharges.clust, parameters = T)
train$TotalCharges <-
factor(TotalCharges.clust$classification)
table(factor(TotalCharges.clust$classification),train$
Churn)%>%chisq.test()

p24+p25

### 对测试集连续变量分箱
testCON <- test %>%
  dplyr::select(where(is.numeric),Churn)

set.seed(2020111142)
tenure.test.clust <-
Mclust(testCON$tenure,G=4,modelNames = 'V')
#(v,4)
summary(tenure.test.clust, parameters = T)
test$tenure <- factor(tenure.test.clust$classification)
table(factor(tenure.test.clust$classification),test$Chu
rn)%>%chisq.test()

MonthlyCharges.test.clust <-
Mclust(testCON$MonthlyCharges,G=4,modelName
s = 'V')  #(v,4)
summary(MonthlyCharges.test.clust, parameters = T)
test$MonthlyCharges <-
factor(MonthlyCharges.test.clust$classification)
table(factor(MonthlyCharges.test.clust$classification
),test$Churn)%>%chisq.test()

TotalCharges.test.clust <-
Mclust(testCON$TotalCharges,G=4,modelNames =
'V')  #(v,4)
summary(TotalCharges.test.clust, parameters = T)
test$TotalCharges <-
factor(TotalCharges.test.clust$classification)
table(factor(TotalCharges.test.clust$classification),te
st$Churn)%>%chisq.test()

#给数据框 WOE 编码
#输入：
#data：数据框
#y:因变量列名
#test：测试集数据框
#输出： WOE 编码后的数据框和 IV 值
#若有 test 则再返回 test 的编码结果
#注： test 和 data 的列名需要完全相同
WOEML <- function(data,y,test=NULL){
  flag <- 1
  if(is.null(test)){
    flag<- 0
    test <- data
  }
  result <- data%>%dplyr::select(eval(y))%>%unlist
  data <- data%>%dplyr::select(-eval(y))
  retest <- test%>%dplyr::select(eval(y))%>%unlist
  test <- test%>%dplyr::select(-eval(y))
  name <- names(data)
  IV <- rep(0,length(name))
  names(IV) <- name
  N0 <- table(result)["0"]
  N1 <- table(result)["1"]
  for(ii in 1:length(name)){
    M0 <-
tapply(unlist(result),list(data[,ii],unlist(result)),lengt
h)%>%as.matrix
    for(jj in 1:nrow(M0)){
      WOE <- log((M0[jj,"1"]/M0[jj,"0"])/(N1/N0))
      IV[ii] <- IV[ii]+(M0[jj,"1"]/N1-
M0[jj,"0"]/N0)*WOE

levels(data[,ii])[which(levels(data[,ii])==rownames(
M0)[jj])] <- WOE

levels(test[,ii])[which(levels(test[,ii])==rownames(M
0)[jj])] <- WOE
    }
    data[,ii] <- as.numeric(as.character(data[,ii]))
    test[,ii] <- as.numeric(as.character(test[,ii]))
  }
  data <- cbind(data,y=result)
  test <- cbind(test,y=retest)
  IV <- sort(IV)
  if(flag){
    return(list(data,IV,test))
  }else{
    return(list(data,IV))
  }
}
```

```
train <- as.data.frame(train)
test <- as.data.frame(test)
WOE <- WOEML(train,y="Churn",test = test)
train.WOE <- WOE[[1]]
test.WOE <- WOE[[3]]
WOE[[2]]

IV.frame <- data.frame(coef = names(WOE[[2]]),value = WOE[[2]])
IV.frame <- IV.frame[order(IV.frame$value),]
IV.frame <- IV.frame[-c(15,11,6),]
IV.frame$coef <- factor(IV.frame$coef,levels = IV.frame$coef)
ggplot(data = IV.frame,aes(x = value, y = coef)) +
  geom_bar(stat = 'identity',
           fill = ifelse(IV.frame$value>0.3,ifelse(IV.frame$value>=0.5,'#F8766D','#00BFC4'),'grey'), # 根据 y 值的正负设置颜色
           width = 0.9)

# 相关性矩阵
ggcorr(train.WOE[-19],label = T,digits = 2,hjust=0.8)
train.WOE %>%

dplyr::select(tenure,InternetService,StreamingTV,StreamingMovies,Contract,MonthlyCharges,TotalCharges)%>%
  ggcorr(label = T,digits = 2,hjust=0.8)

# 建模
# all train
train.WOE.pro <- train.WOE %>%
  select(-MonthlyCharges) # 高度相关 保留其一
fit.logit <- glm(y~.,train.WOE.pro,family = binomial(link = "logit"))
summary(fit.logit)
train.WOE.pro <- train.WOE.pro %>%
  select(-Partner,-tenure_15_MC_75) # 删除系数小于 0 的
fit.logit <- glm(y~.,train.WOE.pro,family = binomial(link = "logit"))
summary(fit.logit)

train.WOE.pro.fac <- train.WOE.pro%>%
  mutate_if(is.numeric,factor,ordered=F)
summary(train.WOE.pro.fac)
summary(train)
# 确定最佳阈值
pred.train.logit.p <- predict(fit.logit, train.WOE.pro, type = 'response')
thresholds <- seq(0,1,by=0.01)
acc.test <- numeric()
accuracy1 <- NULL
accuracy2 <- NULL
for(i in 1:length(thresholds)){
  pred.train.logit <- ifelse(pred.train.logit.p<thresholds[i],"0","1")
  accuracy1 <- confusionMatrix(factor(pred.train.logit,levels = c("0","1")),train.WOE.pro$y)
  accuracy2[i] <- accuracy1$overall[1]
}
acc <- data.frame(p=thresholds,cnt=accuracy2)
opt.p <- subset(acc,cnt==max(cnt))[1,]
sub <- paste("num of parameter:",opt.p$p," accuracy:", opt.p$cnt)
sub

hchart(acc,'line',hcaes(p,cnt))%>%
  hc_title(text='Accuracy With Threshold(logit_WOE)')%>%
  hc_subtitle(text=sub)%>%
  hc_add_theme(hc_theme_google())%>%
  hc_xAxis(title=list(text = 'Number of Parameters'))%>%
  hc_yAxis(title=list(text = 'Accuracy'))

# 使用最优阈值分类
test.WOE.pro <- test.WOE %>%
  select(-Partner,-MonthlyCharges,-tenure_15_MC_75)
pred.logit.p <- predict(fit.logit,newdata = test.WOE.pro,type = 'response')
pred.logit <- ifelse(pred.logit.p < opt.p$p,'0','1')
confusionMatrix(data = factor(pred.logit,levels = c('0','1')),reference = test.WOE.pro$y)

roc.logit <- roc(test.WOE.pro$y,pred.logit.p,quiet=T)
plot(roc.logit,print.auc=T,auc.polygon=T,grid=c(0.1,0.2),max.auc.polygon=T,auc.polygon.col='skyblue',print.thres=T)
```

```
### SVM
# 划分数据集
set.seed(2020111142)
ind <- sample(1:nrow(data),size = 0.7*nrow(data))
train <- data[ind,]
test <- data[-ind,]

# 验证集
ind.val <- sample(1:nrow(train),size = 0.7*nrow(train))
val.train <- train[ind.val,]
val <- train[-ind.val,]

# 调参
gamma <- 2^(-5:5)
cost <- 2^(-5:5)
parms <- expand.grid(cost=cost,gamma=gamma)

acc.test <- numeric()
accuracy1 <- NULL
accuracy2 <- NULL

for (i in 1:NROW(parms)) {
  set.seed(2020111142)
  learn.svm <- svm(Churn~.,data = val.train,kernel = "radial",gamma=parms$gamma[i],cost=parms$cost[i])
  pre.svm <- predict(learn.svm, val[,-17])
  accuracy1 <- confusionMatrix(pre.svm,val$Churn)
  accuracy2[i] <- accuracy1$overall[1]
}

acc.test <- data.frame(p=seq(1,NROW(parms)),cnt=accuracy2)
opt.p <- subset(acc.test,cnt==max(cnt))[1,]
sub <- paste("num of parameter:",opt.p$p," accuracy:", opt.p$cnt)
sub

hchart(acc.test,'line',hcaes(p,cnt))%>%
  hc_title(text='Accuracy With Parameters(SVM)')%>%
  hc_subtitle(text=sub)%>%
  hc_add_theme(hc_theme_google())%>%
  hc_xAxis(title=list(text = 'Number of Parameters'))%>%
  hc_yAxis(title=list(text = 'Accuracy'))
parms$cost[opt.p$p]
parms$gamma[opt.p$p]

# 训练
learn.svm <- svm(Churn~.,train,cost=parms$cost[opt.p$p],gamma=parms$gamma[opt.p$p],probability = T,scale = T)
# 测试
pre.svm <- predict(learn.svm,test[,-17])
confusionMatrix(pre.svm,test$Churn)
pre.svm.p <- as.numeric(predict(learn.svm,newdata=test[,-17],probability=T,type='prob'))
roc.svm <- roc(test$Churn,pre.svm.p,quiet=T)
plot(roc.svm,print.auc=T,auc.polygon=T,grid=c(0.1,0.2),max.auc.polygon=T,auc.polygon.col='skyblue',print.thres=T)

### 随机森林
# 划分数据集
set.seed(2020111142)
ind <- sample(1:nrow(data),size = 0.7*nrow(data))
train <- data[ind,]
test <- data[-ind,]

train <- as.data.frame(train)

# 验证集
ind.val <- sample(1:nrow(train),size = 0.7*nrow(train))
val.train <- train[ind.val,]
val <- train[-ind.val,]

# 调参
acc.test <- numeric()
accuracy1 <- NULL
accuracy2 <- NULL
for(i in 1:16){
  set.seed(2020111142)
  rf.train<-randomForest(Churn~.,data=val.train,mtry=i,ntree=1000)
  rf.pred <- predict(rf.train, val[,-17])
  accuracy1 <- confusionMatrix(rf.pred,val$Churn)
```

```
  accuracy2[i] <- accuracy1$overall[1]
}
acc.test <- data.frame(p=1:16,cnt=accuracy2)
opt.p <- subset(acc.test,cnt==max(cnt))[1,]
sub <- paste("num of parameter:",opt.p$p," accuracy:", opt.p$cnt)
sub

hchart(acc.test,'line',hcaes(p,cnt))%>%
  hc_title(text='Accuracy With mtry(randomforest)')%>%
  hc_subtitle(text=sub)%>%
  hc_add_theme(hc_theme_google())%>%
  hc_xAxis(title=list(text = 'Number of Parameters'))%>%
  hc_yAxis(title=list(text = 'Accuracy'))

fit.rf <- randomForest(Churn~.,data = train,mtry=2,ntree=1000,proximity=T,importance=T)
plot(fit.rf,lwd=3,main = "Random Forest (Error Rate vs. Number of Trees)")
varImpPlot(fit.rf)

pred.rf <- predict(fit.rf,test[,-17])
confusionMatrix(pred.rf,test$Churn)

pre.rf.p <- predict(fit.rf,test[,-17],type="prob")[,1]
roc.rf <- roc(test$Churn,pre.rf.p,quiet=T)
plot(roc.rf,print.auc=T,auc.polygon=T,grid=c(0.1,0.2),max.auc.polygon=T,auc.polygon.col='skyblue',print.thres=T)

###xgboost
# 划分数据集
set.seed(2020111142)
ind <- sample(1:nrow(data),size = 0.7*nrow(data))
train <- data[ind,]
test <- data[-ind,]
train$Churn <- as.character(train$Churn)
train$Churn <- as.numeric(train$Churn)
test$Churn <- as.character(test$Churn)
test$Churn <- as.numeric(test$Churn)
####训练集的数据预处理
# 将自变量转化为矩阵
traindata1 <- data.matrix(train[,-17])
# 利用 Matrix 函数,将 sparse 参数设置为 TRUE,转化为稀疏矩阵
traindata2 <- Matrix(traindata1,sparse=T)
traindata3 <- train$Churn
# 将自变量和因变量拼接为 list
traindata4 <- list(data=traindata2,label=traindata3)
# 构造模型需要的 xgb.DMatrix 对象,处理对象为稀疏矩阵
dtrain <- xgb.DMatrix(data = traindata4$data, label = traindata4$label)


testdata1 <- data.matrix(test[,-17])
testdata2 <- Matrix(testdata1,sparse=T)
testdata3 <- test$Churn
testdata4 <- list(data=testdata2,label=testdata3)
dtest <- xgb.DMatrix(data = testdata4$data, label = testdata4$label)


xgb.fit <- xgboost(data = dtrain,max_depth=2, eta=0.5, objective='binary:logistic', nround=15)
imp.xgb <- xgb.importance(colnames(data[,-17]),model=xgb.fit)
xgb.plot.importance(imp.xgb)

xgb.imp.frame <- data.frame(coef = imp.xgb$Feature,value = imp.xgb$Importance)
xgb.imp.frame <- xgb.imp.frame[order(xgb.imp.frame$value),]
xgb.imp.frame$coef <- factor(xgb.imp.frame$coef,levels = xgb.imp.frame$coef)
ggplot(data = xgb.imp.frame,aes(x = value, y = coef)) +
  geom_bar(stat = 'identity',
           fill = ifelse(xgb.imp.frame$value>0.1,'#F8766D','grey'),
           width = 0.9)

xgb.cov.frame <- data.frame(coef = imp.xgb$Feature,value = imp.xgb$Cover)
xgb.cov.frame <- xgb.cov.frame[order(xgb.cov.frame$value),]
xgb.cov.frame$coef <- factor(xgb.cov.frame$coef,levels = xgb.cov.frame$coef)
ggplot(data = xgb.cov.frame,aes(x = value, y = coef)) +
```

```
  geom_bar(stat = 'identity',
           fill =
ifelse(xgb.cov.frame$value>0.1,'#F8766D','grey'),
           width = 0.9)

pred.train.xgb.p <- predict(xgb.fit,dtrain)

# 確定最佳閾值
thresholds <- seq(0,1,by=0.01)
acc.test <- numeric()
accuracy1 <- NULL
accuracy2 <- NULL
for(i in 1:length(thresholds)){
  pred.train.xgb <-
ifelse(pred.train.xgb.p<thresholds[i],"0","1")
  accuracy1 <-
confusionMatrix(factor(pred.train.xgb,levels =
c("0","1")),as.factor(train$Churn))
  accuracy2[i] <- accuracy1$overall[1]
}
acc.test <- data.frame(p=thresholds,cnt=accuracy2)
opt.p <- subset(acc.test,cnt==max(cnt))[1,]
sub <- paste("num of parameter:",opt.p$p,"
accuracy:", opt.p$cnt)
sub

hchart(acc.test,'line',hcaes(p,cnt))%>%
  hc_title(text='Accuracy With
Threshold(xgboost)')%>%
  hc_subtitle(text=sub)%>%
  hc_add_theme(hc_theme_google())%>%
  hc_xAxis(title=list(text = 'Number of
Parameters'))%>%
  hc_yAxis(title=list(text = 'Accuracy'))

pred.test.xgb.p <- predict(xgb.fit,dtest)
pred.xgb <- ifelse(pred.test.xgb.p < opt.p$p,'0','1')
confusionMatrix(data = factor(pred.xgb,levels =
c('0','1')),as.factor(test$Churn))
roc.xgb <-
roc(as.factor(test$Churn),pred.test.xgb.p,quiet=T)
plot(roc.xgb,print.auc=T,auc.polygon=T,grid=c(0.1,0
.2),max.auc.polygon=T,auc.polygon.col='skyblue',pr
int.thres=T)
```