基于大数据平台的

XX 大数据应用场景

模型分析报告

二〇一五年三月

1 概述

1.1 编写目的

本文档是 XXXXX 挖掘模型设计说明书，描述了数据挖掘模型的完整建设过程，包括业务目标定义、数据理解、数据预处理、模型构建、模型评估、模型应用等主要建设过程。

本文档用于指导数据挖掘相关人员搭建模型分析与部署，包括：

- 数据挖掘宽表设计；
- 数据预处理；
- 模型的变量设计；
- 模型评估及应用。

1.2 术语定义

序号	名词	相关解释
1	客户价值	指客户为企业提供的价值，即从企业角度出发，根据客户消费行为和消费特征等变量测量出客户能够为企业创造的价值，该客户价值衡量了客户对于企业的相对重要性，是企业进行差异化服务的重要标准

2 CRISP-DM 数据挖掘实施方法论

CRISP-DM(Cross-Industry Standard Process For Data Mining)指的是跨行业数据挖掘标准过程，包括商业理解(Business

Understanding)、数据理解(Data Understanding)、数据准备(Data Preparation)、建立模型(Modeling)、模型评估(Evaluation)、模型发布(Deployment)等部分内容。

1.商业理解（business understanding）

在这第一个阶段我们必须从商业的角度了解项目的要求和最终目的是什么，并将这些目的与数据挖掘的定义以及结果结合起来.

主要工作包括：确定商业目标，发现影响结果的重要因素，从商业角度描绘客户的首要目标，评估形势，查找所有的资源、局限、设想以及在确定数据分析目标和项目方案时考虑到的各种其他的因素，包括风险和意外、相关术语、成本和收益等等，接下来确定数据挖掘的目标，制定项目计划。

2.数据理解（data understanding）

数据理解阶段开始于数据的收集工作。接下来就是熟悉数据的工作，具体如：检测数据的量，对数据有初步的理解，探测数据中比较有趣的数据子集，进而形成对潜在信息的假设。收集原始数据，对数据进行装载，描绘数据，并且探索数据特征，进行简单的特征统计，检验数据的质量，包括数据的完整性和正确性，缺失值的填补等。

3.数据准备（data preparation）

数据准备阶段涵盖了从原始粗糙数据中构建最终数据集(将作为建模工具的分析对象)的全部工作。数据准备工作有可能被实施多次，而且其实施顺序并不是预先规定好的。这一阶段的任务主要包括：制表，记录，数据变量的选择和转换，以及为适应建模工具而进行的数据清理等等。

根据与挖掘目标的相关性，数据质量以及技术限制，选择作为分析使用的数据，并进一步对数据进行清理转换，构造衍生变量，整合数据，并根据工具的要求，格式化数据。

4.建立模型（modeling）

在这一阶段，各种各样的建模方法将被加以选择和使用，通过建造，评估模型将其参数将被校准为最为理想的值。比较典型的是，对于同一个数据挖掘的问题类型，可以有多种方法选择使用。如果有多重技术要使用，那么在这一任务中，对于每一个要使用的技术要分别对待。一些建模方法对数据的形式有具体的要求，因此，在这一阶段，重新回到数据准备阶段执行某些任务有时是非常必要的。

5.模型评估（evaluation）

从数据分析的角度考虑，在这一阶段中，已经建立了一个或多个高质量的模型。但在进行最终的模型部署之前，更加彻底的评估模型，回顾在构建模型过程中所执行的每一个步骤，是非常重要的，这样可以确保这些模型是否达到了企业的目标。一个关键的评价指标就是看，是否仍然有一些重要的企业问题还没有被充分地加以注意和考虑。在这一阶段结束之时，有关数据挖掘结果的使用应达成一致的决定。

6.模型部署（deployment）

部署,即将其发现的结果以及过程组织成为可读文本形式.模型的创建并不是项目的最终目的。尽管建模是为了增加更多有关于数据的信息，但这些信息仍然需要以一种客户能够使用的方式被组织和呈

现。这经常涉及到一个组织在处理某些决策过程中，如在决定有关网页的实时人员或者营销数据库的重复得分时，拥有一个“活”的模型。

根据需求的不同，部署阶段可以是仅仅像写一份报告那样简单，也可以像在企业中进行可重复的数据挖掘程序那样复杂。在许多案例中，往往是客户而不是数据分析师来执行部署阶段。然而，尽管数据分析师不需要处理部署阶段的工作，对于客户而言，预先了解需要执行的活动从而正确的使用已构建的模型是非常重要的。

3 XXX（模型名称）

3.1 业务目标

通过客户特征分析，对客户价值进行细分分析，实现动态识别客户价值；分类识别客户违约风险，促进用电检查、电费回收的管理水平提升，降低电力企业的经营风险，提高企业运作效率。

3.2 数据理解

3.2.1 输入业务信息

业务信息类	内容
用户基本信息	统计月份
	客户名称
	客户编号
	客户年龄
	客户身份
	客户状态
	所属地市
	所属行业
	客户类别
	用电类别
	是否存在预购电装置
	城农网标识
	预付费标识
	重要客户标识
	高可靠性标识
	信用等级
	合同容量
	供电电压
	负荷类型
	总用电量等级划分
	抄表周期

3.2.2 输出业务信息

业务信息类	内容
用户基本信息	统计月份
	客户名称

	客户编号
	客户年龄
	客户身份
	客户状态
	所属地市
	所属行业
	客户类别
	用电类别
	是否存在预购电装置
	城农网标识
	预付费标识
	重要客户标识
	高可靠性标识
	信用等级
	合同容量
	供电电压
	负荷类型
	总用电量等级划分
	抄表周期

3.3 数据预处理

1. 缺失值处理

若用户数据整点负荷值连续多个为空（可设置连续几个为空时进行数据过滤），则将该数据过滤掉。根据输入表中 data_whole_flag_24 字段中的 X 表示没有相应采集点、0 和 9 都表示数据异常。

未过滤的数据，对缺失值处理如下：

方法一：前一时刻值填充法：利用前一时刻的负荷值填充到下一时刻的空缺值中。例如：对日/月负荷的数据利用前一天/前一个月的数据进行替换空值，对于日分时负荷利用前一个小时的数据进行替换空值。

方法二：五（七）点平滑填充法（除此之外还包括平均值平滑、边界值平滑、中值平滑）：利用前后五点（七点）的负荷值得均值填充到空缺值中。例如：对日/月负荷的数据利用前后五天/月的数据进行替换空值，对于日分时负荷利用前后五个小时的数据进行替换空值。

方法三：可以利用多项式回归、贝叶斯形式化方法工具或判定树归纳等确定空缺值。这类方法依靠现有的数据信息来推测空缺值，使空缺值有更大的机会保持与其他属性之间的联系。

在归一化处理前，分别采用三种方法进行处理。

2. 归一化处理

负荷数据归一化处理计算公式：

$$\overline{P}_{ob,i,j}=\frac{P_{ob,i,j}-P_{ob,i,Min}}{P_{ob,i,Max}-P_{ob,i,Min}}$$

其中，ob 代表用户编号，i∈［1～31］，表示日；j∈［0～23］，表示小时；，表示该用户每日负荷的最大值和最小值；表示客户的整点负荷；表示归一化后的用户某日某时的负荷，最大值为 1。

3.4 模型构建

3.4.1 模型构建思路

首先选取有可能影响迎峰度夏期间配变发生重过载的信息数据，如配变的历史负荷数据、配变所属区域数据、设备信息、客户信息数据、气温数据等；然后，综合考虑短期负荷周期性波动，负荷影响“近大远小”的规律，气温、日照等气象因素对迎峰度夏期间配变负荷的影响，设计相关的特征量。最后，构建短期预警模型，对短期重过载现象进行预警分析。

3.4.2 算法选择

短期预警模型的建立对电力系统近期输变电建设、运行和计划都非常重要。短期负荷除具有明显的周期性外，还受到各种环境因素的影响，如天气因素、季节变换、电力市场、重大事件等，使得负荷的时间序列变化呈现出非平稳的随机过程。由于短期预警的不准确性和条件性，因此要对负荷在各种情况下可能的发展状况进行预警有一定难度。短期预警的建模算法有多种，它们也都有各自的特点和适用条件。在分析短期负荷特点和影响因素的基础上，综合考虑短期预警算法的特点，分析归纳出三种比较适合可行的算法，其分别是 Logistic 回归模型、多元回归模型和时间序列模型，下面分别对这三种方法的理论进行介绍。

3.4.2.1 算法方法介绍

3.4.2.1.1 Logistic

传统线性回归模型在实际定量分析中，受到多种条件的限制。例如，当因变量是一个分类变量而不是一个连续变量，线性回归模型不再适用。在分析分类变量时，通常采用的统计方法是对数线性模型（Log-linear-model）。在本期研究的课题是对配变是否发生重过载现象进行短期预警。配变是否发生过重过载，可分为“发生”和“不发生”两类，也被称为二分变量。当对数模型中的一个二分变量被当作因变量时，对数线性模型即成为 Logistic 回归模型。Logistic 回归分析为解决这种因变量为定性变量的问题提供了有效的分析工具。

设Y为一个随机变量，且服从两点分布。当事件发生时，Y的取值为 1，当事件没有发生时，Y的取值为 0。

假设影响实验结果的自变量为$X=(X_1,X_2,\cdots,X_p)$，则在给定X的条件下，$Y=1$的概率为：

$$P(Y=1|X)=\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}$$ （式 4.1.1）

公式 4.1.1 就是 Logistic 函数，它具有 S 分布，如图 4.1.2:

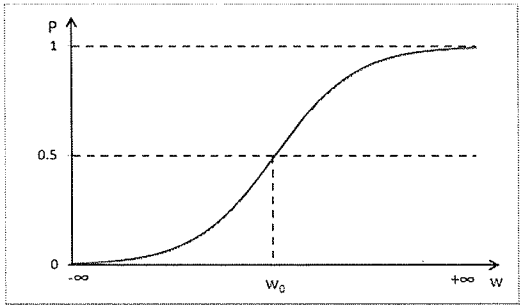

图 4.1.2 Logistic 函数的 S 曲线图

令 $w=\beta_0+\sum_{i=1}^{p}\beta_i x_i$，可以看出，当 w 趋于负无穷大时，$P(Y=1|X)=\lim_{w\to-\infty}\frac{e^w}{1+e^w}=0$，当 w 趋于正无穷大时，$P(Y=1|X)=\lim_{w\to+\infty}\frac{e^w}{1+e^w}=1$。正如图 2.1.1 所示，无论 w 取任何值，Logistic 函数的取值范围均在 0-1 之间变动。Logistic 函数的这一性质保障了由 Logistic 模型估计的概率决不会大于 1 或小于 0。同时这个函数的形状对于研究概率也很适合，当 w 从负无穷开始向右移动时，函数值先是很缓慢地增加，在接近 w_0 时开始迅速增加，之后增加的速度又开始逐渐减缓，最后当 w 趋于正无穷大时，函数值趋于 1。Logistic 函数的 S 曲线表明，w 的作用对于案例发生某一事件的可能性是变化的，在 w 值很小时其作用也很小，然后在中间阶段对应的可能性增加很快，但是在 w 值增加到一定程度以后，可能性就保持在几乎不变的水平，这说明，w 在 $P(Y=1|X)$[1]接近于 0 或 1 时的作用要小于当 $P(Y=1|X)$ 处于中间阶段时的作用。

由 Logistic 回归的定义可知，事件发生的概率为：

[1] $P(Y=1|X)$ 指事件发生，及 Y=1 时的概率大小

$$\pi=P(Y=1|X)=\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}\quad（式 4.1.2）$$

所以，事件不发生的概率为：

$$\begin{aligned}1-\pi&=1-P(Y=1|X)\\&=1-\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}\\&=\frac{1}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}\end{aligned}\quad（式 4.1.3）$$

因此，由公式 4.1.2 和公式 4.1.3 可以得出：

$$\frac{\pi}{1-\pi}=e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}\quad（式 4.1.4）$$

我们称比率 $\frac{\pi}{1-\pi}$ 为事件的优势比，对上式进行对数变换，则有：

$$g(x_1,x_2,\cdots,x_n)=\log(\frac{\pi}{1-\pi})=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p\quad（式 4.1.5）$$

优势比的对数称为 Logit。Logit 变换产生了参数为 $\beta_0,\beta_1,\beta_2,\cdots,\beta_p$ 的一个线性函数，拟合 Logistic 回归模型的参数问题转换为拟合线性模型的参数。通常采用极大似然法（后文简称 MLE）来估计参数。下面介绍 Logistic 回归模型的假设前提、适用条件及验证方法。

Logistic 回归模型也有其假设前提适用条件及验证方法。Logistic 回归模型估计的假设条件与 OLS 线性回归分析的假设条件有相似之处，也有所区别。相似的地方包括：首先，数据必须是总体数据或者来自随机样本。第二、因变量 y_i 被假设为 K 个自变量 $x_{ki}\ (k=1,2,\cdots,K)$ 的函数。第三，Logistic 回归也对多元共线

性敏感，自变量之间存在的多元共线性会导致标准误的膨胀。

Logistic 回归模型还有一些与 OLS 回归不同的假设。第一，Logistic回归的因变量y_i是二分变量，这个变量只能取值0或1。研究的兴趣在于事件发生的条件概率，即$P(y_i=1|x_{ki})$。第二，Logistic回归中因变量和各自变量之间的关系是非线性的。第三，在 OLS 回归中要假设相同分布性或称方差不变，类似的假设在 Logistic 回归中却不需要。最后，Logistic 回归也没有关于自变量分布的假设条件。各自变量可以是连续变量，也可以是离散变量，还可以是虚拟变量。并且，也不需要假设它们之间存在多元正态分布。

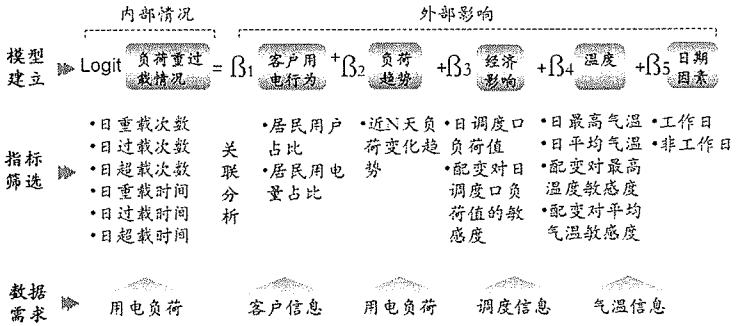

图 4.1.3 短期的 Logistic 模型

在短期预警中应用此模型的建模框架如图 4.1.3 所示，通过输入客户、用电负荷、调度和气温等维度的特征信息建立 Logistic 模型，然后用模型对配变未来发生重过载的概率进行预警。

3.4.2.1.2 多元回归

多元回归分析预测法，是指通过对两上或两个以上的自变量与一个因变量的相关分析，建立预测模型进行预测的方法。其模型的数学表达式如下：

设y是一个可观测的随机变量，它受到p个非随机因素x_1，$x_2,\ldots,x_p$和随机因素ε的影响，若y与$x_1,x_2,\ldots,x_p$有如下线性关系：

$$y=\beta_0+\beta_1x_1+\cdots+\beta_px_p+\varepsilon \quad （式 4.1.6）$$

其中β_0，$,\ldots,\beta_p$是$p+1$个未知参数，ε是不可测的随机误差，且通常假定$\varepsilon\sim N(0,\sigma^2)$.我们称式（4.1.6）为多元线性回归模型.称$y$为被解释变量（因变量），$x_i(i=1,2,\cdots,p)$为解释变量（自变量）。

对式 4.1.6 取期望得

$$E(y)=\beta_0+\beta_1x_1+\cdots+\beta_px_p \quad （式 4.1.7）$$

式 4.1.7 为理论回归方程。

对于一个实际问题，要建立多元回归方程，首先要估计出未知参数β_0,β_1，$\ldots,\beta_p$，为此我们要进行n次独立观测，得到n组样本数据$(x_{i1},x_{i2},\cdots,x_{ip};y_i)$，$i=1,2,\cdots,n$。然后对未知参数进行估计，与一元线性回归时的一样，多元线性回归方程中的未知参数$\beta_0,\beta_1,\cdots,\beta_p$仍然可用最小二乘法来估计。

给定因变量y与$x_1,x_2,\ldots,x_p$的n组观测值，利用前述方法确定线性回归方程是否有意义，还有待于显著性检验。主要包括回归方程显著性的F检验、回归系数的t检验以及衡量回归拟合程度的拟合优度检验。

以上便是多元回归模型算法，下面具体举例来说明多元回归模型的具体结构。

$$y_t = \beta_0 + \beta_1 \times Temp + \beta_2 \times Temp^2 + \beta_3 \times Load + \sum_{i=1}^{7}\sum_{j=6}^{9} s(i,j,t)\cdot c_{ij} + d\cdot t \quad （式4.1.8）$$

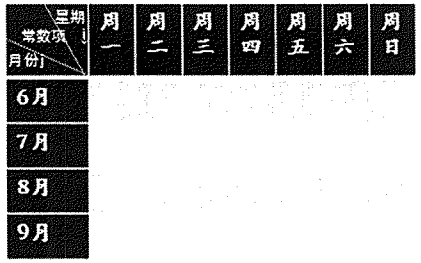

图 4.1.4

在短期总应用多元回归建模大体结构如式 4.1.8，式中，t 表示日期，y 表示日电量，Temp 表示平均温度，β_0表示回归模型的中常数项，Temp 表示温度信息，Load 表示符合信息，$d\cdot t$ 表示电量随时间的线性增长趋势。符号函数$s(i,j,t)$来表征不同星期类型、不同月份的影响，如图 4.1.4 所示，当且仅当第 t 日的星期恰好等于 i，月份恰好等于 j 时，$s(i,j,t)$才取值为 1，否则函数值一律为 0。c_ij表示星期类型为 i，月份为 j 时的回归常数项。

3.4.2.1.3 时间序列

时间序列是指以时间顺序形态出现的一连串观测值的集合，或者更确切的说，是对某动态系统随时间连续观察所产生的有顺序的观测值的集合。在统计研究中，常用按时间序列排列的一组随机变量（…，$X_1, X_2, \cdots, X_t, \cdots$）来表示一个随机事件的时间序列，简记为$\{X_t,\ t \in T\}$或$\{X_t\}$。用$x_1, x_2, \cdots, x_n$表示该随机序列的 n 个有序观察值，称之为序列长度为 n 的观察值序列。一个观察值序列属于随机过程的一次样本实现。随机时间序列的一个最基本特征就是相邻两个数据之间有相互依赖性，即：两个随机数据呈现一定的相关。时间序列分析就是依据不同时刻变量的相关关系进行分析，生成随机动态模型来揭示其相关结构并进行预测。模型的数学表达式如下：

若有平稳零均值随机序列$\{X_t\}$及白噪声序列$\{\alpha_t\}$[2]满足

$$X_t - \phi X_{t-1} - \cdots - \phi_p X_{t-p} = \alpha_t - \theta_1 \alpha_{t-1} - \cdots - \theta_t \alpha_{t-q} \quad （式 4.1.9）$$

引入延迟算子 B，记

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$
$$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \quad （式 4.1.10）$$

则式(4.1.9) 又可写为

$$X_t \Phi(B) = \alpha_t \Theta(B) \quad （式 4.1.11）$$

若 $\Phi(B) = 0$ 与 $\Theta(B) = 0$ 的根都在单位圆外，则上面的模型即为 ARMA 模型[5]。它是时间序列法的一般形式，可视为一个单入单出的线性系统。当将 ARMA 模型用于预报时，$\{\alpha_t\}$就是残差序列。

[2]唐鸿龄，等. 应用概率[M]. 南京：南京工学院出版社, 1988.

表 4.1.3 ARMA（p,q）模型的自相关系数和偏自相关系数特征

模型	相关系数	偏自相关系数
AR（p）	拖尾	P 阶截尾
MA（q）	q 阶截尾	拖尾
ARMA（p,q）	拖尾	拖尾

如表 4.1.3，作为 ARMA 的特例,当 q=0 时称为 AR(p)(p 阶自回归)模型。其特点为:偏自相关函数具有截尾性,自相关函数具有拖尾性;当 p=0 时称为 MA(q)(q 阶滑动平均)模型,其自相关函数具有截尾性,偏相关函数具有拖尾性;对非平稳序列作 d 阶差分后再拟合 ARMA 模型则称为 ARIMA(p,d,q)模型。

3.4.2.2 算法优选

表 4.1.4 模型对比分析

	Logistic	多元回归	时间序列
输入	各个特征量信息	各个特征量信息	历史负荷信息
输出	每台配变未来一周重过载概率清单	每台配变每天负荷值	每台配变每个小时负荷值
可解释性	高	中	低
预测精度	中	中	高
建模个数	1	建模个数等于配变个数	建模个数等于配变个数
实现难度	中	中	高

根据表 4.1.4 中，综合评估 3 种算法在输入、输出、可解释性、预测精度、建模个数和实现难度上的优劣。在本次建模中，采用 Logistic 算法作为建模算法的首选方案。多元回归和时间序列作为备选方案。

3.5 模型评估

3.5.1 模型效果评价方法介绍

短期预警模型预警结果的好坏关系到基于其结果做出决策的损失大小，好的、稳定的、与实际拟合度高的模型能为决策层提供可靠的参考依据，因此对模型预警能力的评价至关重要，这就要求有一系列科学合理的评价指标从不同的需求考虑对模型预警能力做出量化的衡量，由于不同算法建立的模型输出结果不同，因此针对不同模型也有不同的评价标准，下面介绍几种评价标准并给出它们适用的算法。

3.5.1.1 查全率、查准率

表 4.1.5 混淆矩阵

实际 / 预测	不发生	发生	查准率
不发生	202	3	
发生	2	5	71.43%
查全率		62.50%	97.64%

混淆矩阵（confusion matrix）是可视化工具，每一列代表了实际测得信息，每一行代表了预测的分类信息，查全率和查准率正是基于混淆矩阵上计算得出的。

查全率是在实际发生的范畴中讨论预测的准确性，它是衡量预测模型正确预测实际发生的能力，其计算公式为:（预测发生且实际发生的台变数/实际发生的总台变数）*100%。如表 4.1.5，查全率=5/(3+5)*100%=62.50%。

查准率是在预测发生的范畴中讨论预测的准确性，它是对预测模型预测发生中预测正确的肯定，其计算公式为：（预测发生且实际发生的台变数/预测发生的总台变数）*100%。如表 4.1.5，查准率=5/(2+5)*100%=71.43%。

由于查全率和查准率是根据混淆矩阵中分类信息计算出的指标，在三种模型中只有 Logistic 模型能输出混淆矩阵中的分类信息，故查全率和查准率只适合 Logistic 模型的评价。

3.5.1.2 准确率

准确率又称“精度”、“正确率”，用来衡量预测值与实际值的拟合程度，其计算公式为：（|预测值-实际值|/实际值）*100%。

由于准确率的计算公式适用的是连续变量而不是分类变量，而只有多元回归模型和时间序列模型的输出结果是连续变量，故该评价指标适用多元回归模型和时间序列模型。

3.5.2 模型优化方法介绍

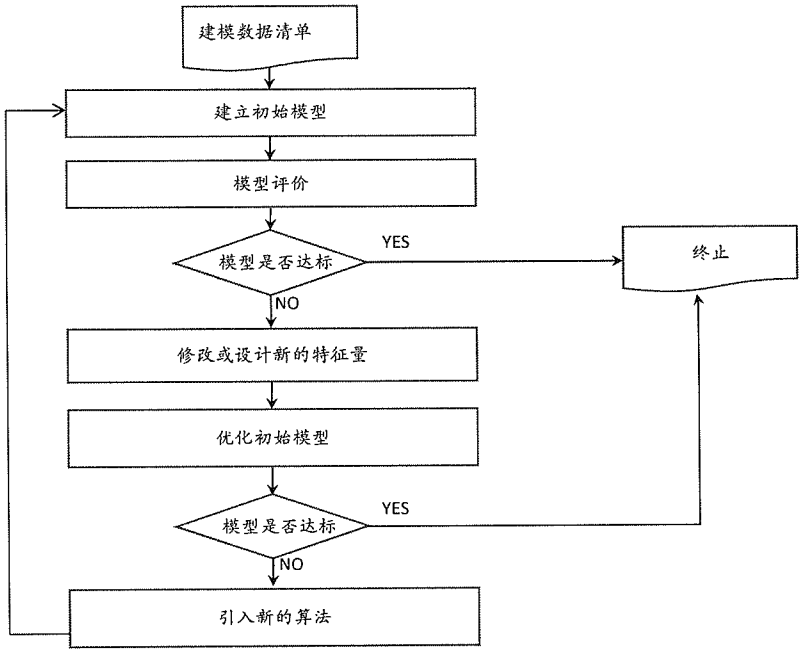

图 4.1.5 模型优化流程

模型的优化流程如图 4.1.5，首先获取建模所需的数据清单，然后建立初始模型，通过对模型的查全率、查准率和正确率的评价是否达到预警目的的要求，如果达到模型的预警要求则终止，此时初始模型即为最终模型，如果达不到预警目的要求，则通过修改或设计新的特征量优化初始模型，然后重新评价模型是否达到预警要求，是则优化后的模型为最终模型，否则考其它算法重新建模。直至模型达到预警要求为止。

3.6 模型应用

根据不同算法特点，其输出的结果也不尽相同，下面对不同算法建立的预警模型的预期输出结果进行介绍。

概率排名	配变编号	配变名称	未来一周		是否重过载	
			重载概率	过载概率	重载	过载
Top1			xx%	xx%		
Top2			xx%	xx%		
…			…	…		
Top4			xx%	xx%		

图 4.1.6 Logistic 预警模型输出清单

由于 Logistic 模型的输出结果是每台配变的概率，因此其输出的清单是一张配变可能发生重过载的概率表，如图 4.1.6 所示，结果主要包含以下字段信息：概率排名、配变编号、配变名称、未来一周重过载概率和是否重过载的判断。

概率排名	配变编号	配变名称	第1天			第2天			第N天		
			1h	…	24h	1h	…	24h	1h	…	24h
Top1			xx	xx	xx	xx	xx	xx	xx	xx	xx
Top2			xx	xx	xx	xx	xx	xx	xx	xx	xx
…			…	…	…	…	…	…	…	…	…
Top4			xx	xx	xx	xx	xx	xx	xx	xx	xx

注：短期预警模型清单适用条件：预警时间范围第 N 天根据天气预报范围、调度区负荷预测值时间范围以及预测精度而定，默认为 7 天。

图 4.1.7 多元回归和时间序列模型输出清单

由于多元回归模型和时间序列模型的输出结果是每台配变每个时间单位的具体负荷，因此其输出的清单是一张配变可能承当的负荷程度，如图 4.1.7 所示，结果主要包含以下字段信息：概率排名、配变编号、配变名称、未来 N 天每个时点的配变预测负荷值。

将以上结果做好输出准备以备场景构建使用。

4 附录

4.1 数据需求表

4.2 数据预处理常用方法

1. 数据清理：缺失值处理

- 忽略该记录（元组）
  - 通常在进行分类、描述、聚类等挖掘任务，但是元组缺失类标识时
  - 该方法通常不是最佳的，尤其是缺失数据比例比较大的时候
- 手工填入空缺的值
  - 枯燥、费时，可操作性差，不推荐使用
- 使用一个全局常量填充空缺的值
  - 给定一个固定的属性值，如：未知、不详、Unknow、Null 等
  - 简单，但意义不大

- 使用属性的平均值填充空缺值
  - 简单方便，但是挖掘结果容易产生不精确的结果
- 使用与给定元组同一个类别的所有样本的平均值
  - 分类非常重要，尤其是分类指标的选择，常常需结合业务含义
- 使用最有可能的值进行填充
  - 利用回归、决策树、贝叶斯形式化推导方法
  - 优点是利用属性之间的关系进行推断，保持了属性之间的联系

2. 数据清理：平滑噪音数据

- 分箱（Binning）方法
- 聚类方法
  - 检测并剔除异常点
- 线性回归
  - 利用回归函数进行平滑处理
- 人机结合共同检测
  - 计算机检测出的可疑点，通过人来确认

3. 数据归一化

归一化是一种简化计算的方式，即将有量纲的表达式，经过变换，化为无量纲的表达式，成为纯量，避免具有不同物理意义和量纲的输入变量不能平等使用。在统计学中，归一化的具体作用是归纳统一样本的统计分布性。常用的方法有：零-均值规范化、最小-最大规范化。

4.3 模型分析报告模板（WORD 版本及 PPT 版本）