摘 要

基于车联网数据的驾驶行为安全评分方法研究

随着经济的发展和人们生活需求的不断变化，金融领域发展出的评分卡模型给很多行业都带来便利，例如银行就可以借助评分卡模型规避掉一些信贷风险，帮助企业处于良性的发展状态。在金融领域，学者们借助信贷数据有明确标注的优势，广泛地将机器学习方法和评分卡模型相结合，不断地丰富模型的种类、可适用范围。目前评分卡模型的研究背景多数集中在金融、保险行业。

被大家熟知的汽车领域，尤其是网联汽车领域目前也在迅速发展。研究影响车辆行驶安全的因素是至关重要的，例如情绪、年龄、性别、路况、天气、驾龄等。此外，对车速进行研究，得到侧向摩擦因数与速度的关系，能够对驾驶特征进行分类，为数据的深度挖掘和构建用户画像奠定了基础。

把金融评分卡模型迁移到车联网领域，建立驾驶行为安全评分卡模型是非常具有实践意义的。本文选取某市的车联网数据，建立驾驶行为评分卡，从而对驾驶安全进行评估。通过数据预处理，完成构建评分卡所需的数据准备工作。本文首先建立基于 Logistic 回归的驾驶安全评分卡，给出在 Logistic 回归建模场景下，预判有危险驾驶行为的判定准则。考虑特征选择对于评分卡建模的重要性，本文基于 BP 神经网络的 MIV 特征选择算法，建立 XGBoost 驾驶行为安全评分卡，并给出在该建模场景下的判定准则。

关键词： 车联网数据，BP 神经网络，评分卡

Abstract

Driving Behavior Safety Scoring Method Based on Internet of Vehicles Data

With the development of economy and the continuous change of people's life needs, the scorecard model developed in the financial field has brought convenience to many industries. For example, banks can use the scorecard model to avoid some credit risks and help enterprises in a healthy state of development. In the field of finance, scholars have widely combined machine learning methods and scorecard models with the advantage of clearly labeled credit data, and continuously enriched the types and applicable scope of models. At present, most of the research background of scorecard model focuses on the finance and insurance industry.

The well-known field of vehicles, especially the field of Internet of Vehicles, is also rapidly developing. It is essential to study the factors that affect the safety of vehicle driving, such as emotion, age, gender, road conditions, weather, driving age, etc. In addition, the vehicle speed is studied to obtain the relationship between side friction factor and speed. Then, the driving feature can be classified, which lays a foundation for the depth mining of data and the construction of user portraits.

It is of great practical significance to migrate the financial scorecard model to the field of Internet of Vehicles and establish the driving behavior safety scorecard model, so as to evaluate driving safety. In this thesis, the Internet of Vehicles data in a city are selected to establish driving behavior scorecard. Through data preprocessing, data preparation for the construction of scorecard is completed. This thesis first establishes the driving safety scorecard based on Logistic regression. In the scenario of Logistic regression modeling, the judgment criterion of predicting dangerous driving behavior is given. Consider the importance of feature selection for scorecard modeling. Then the MIV feature selection algorithm based on BP neural network in this thesis, establish XGBoost driving safety scorecard. And the judgment criterion in this modeling scenario is given.

Keywords: Internet of Vehicles data, BP neural network, Scorecard

目　录

第一章　绪论

§1.1　研究背景

随着经济社会的不断发展，数字时代早已到来。1963 年，James H 等人 [1] 将评分卡模型应用到金融行业，从而促进了银行信贷业务的发展和完善，信贷风险也因评分系统的完善逐渐降低。经典的评分模型 [2] 通常被用于银行信贷业务，如信贷业务的申请 [3]、催收 [4]、借贷 [5] 等。评分卡模型使用的样本是有真实标签的二分类数据 [6]，例如将类别划分为诚信和违约。由于企业和个人对信贷业务的需求不断增加，金融行业为了降低信贷的风险，需要对申贷人进行风险评估 [7]，于是评分模型最先在金融行业被大规模使用并不断发展成熟。目前，评分模型主要应用在金融行业，同时也被推广到房地产和保险行业，而在汽车领域，尤其在网联汽车的应用上还有很大的发展空间。

车联网 (Internet of Vehicles，IoV) 是通过现代信息技术将驾驶员、车辆以及路基三者进行信息采集和融合，借助互联网、物联网达到信息互联互通的目的，是汽车领域发展的一项重要成果。“大数据”时代 [8]，车联网的发展使得大量的数据能够被采集和存储，其中包含大量的驾驶行为信息 [9]，例如车速、驾驶员安全带状态、发动机转速等。这些驾驶行为特征构成的数据成为机器学习适合的数据来源之一。随之，利用经典机器学习方法在车联网数据上进行训练，得到高精度的模型。车联网兴起的作用之一，是为改善交通文明，保护人民生命财产安全，能够为出行安全提供一定的保障。考虑到决策树、k 近邻、支持向量机、随机森林等 [10] 分类算法的优势所在，学者们将有标注的驾驶行为数据进行分类预测，并对具有潜在危险的驾驶行为进行筛选，能够在一定程度上对驾驶安全有积极的促进作用。利用这些驾驶数据建立一种合理的打分机制，使得各个驾驶特征都能够给定一个分值，将驾驶行为的优劣通过分数进行量化，是十分必要的。

随着评分卡模型的发展，各行各业越来越重视建立评分系统的必要性。于是，本文在前人的研究基础之上，对驾驶数据进行探索分析，建立基于车联网数据的驾驶行为安全评分卡，多维度分析驾驶行为，并对驾驶行为进行安全评估。

§1.2 研究现状

信用风险最早在国外受到重视并发展，目前为止信用评分体系已经发展的较为成熟。Fair B 和 Isaac E [1] 共同构建了信贷评分系统 Fair Isaac Corporation (FICO)，推动了金融领域的发展进程。Durand D 等人 [11] 利用信贷客户的借贷行为数据进行建模，定量地评估信用风险。Ruddy John A 等人 [12] 利用不同评级机构的数据建立评分，帮助银行业降低投资风险。夏万永 [13] 利用带有真实标签的房地产数据训练模型，验证得到随机森林、向前逐步回归、Boruta 算法等机器学习方法在众多特征选择方法中表现良好，并将评分卡模型应用到房地产领域，为预防和化解信用欺诈风险提供了切实可行建议和方案。刘荣珍 [14] 借助金融公司 31200 条有着真实标签的用户信贷数据，结合 BP 神经网络、网格搜索、贝叶斯优化等机器学习方法构建评分卡模型，评估信贷用户风险，帮助金融机构做出合理的房贷决策。

在汽车相关领域，刘拥华等人 [15] 表明大量道路事故是因车速选择不当或因至少一方先违反交通规则造成的。多数研究表明：路况、天气条件以及交通量 [16]、不同任务需求 [17]、恐惧和焦虑 [18]、心理素质和情绪起伏 [19]、年龄 [20]、性别 [21]、驾龄 [22] 等因素会影响驾驶员实时调整驾驶行为，从而不同程度地影响车辆行驶安全。Rashid Izullah F 等人 [23] 通过让受试者参与虚拟驾车环境，模拟出驾驶数据，研究得出不良的情绪状态，会增加危险驾驶的风险。Tolonen J 等人 [24] 和 Palazón-Bru A 等人 [25] 基于交通事故调查，研究结果显示，单次机动车事故中，未系安全带死亡率比使用安全带的受伤率还高。因此，驾驶安全和生命健康息息相关。Goebelbecker J M 和 Uzgiris S C 等人 [26] 通过实验数据研究表明车速是造驾驶安全的主要因素。Bagdadi O 等人 [27] 研究得出速度越高，轮胎能够承受的侧向摩擦力越小。Eboli L 等人 [28] 通过牛顿第二定律推导出通过速度计算横纵摩擦因数的公式，以摩擦因数为桥梁得到速度和加速度曲线，并将驾驶特征进行分类。在此研究基础上，研究者们结合 K 近邻 [29]、决策树 [30]、SVM [31]、前馈神经网络 [32]、云模型 [33] 等机器学习的方法对驾驶特征进行分析预测。Cheng Runtong 等人 [34] 通过对汽车的车门、车灯、车胎等部位的工作状态进行评估，能够为汽车性能

提供一个报警机制，帮助驾驶员自动检测汽车性能，从而为驾驶安全提供保障。

目前在驾驶数据上的研究，大多数仍表现在对特征进行分类，追求分类达到的精度，对算法性能的提升，体现机器学习算法在分类层面的优势。由于金融领域评分体系发展的比较完善，给整个行业都带来了安全保障，降低了金融风险。因此，把金融评分卡模型迁移到车联网领域，构建驾驶行为安全评分卡模型是十分具有实践意义的。

§1.3 本文结构

本文采用某市三月份的车联网数据，采用数据处理分析技术，对驾驶行为特征进行分析和筛选，建立驾驶行为安全评分模型，本文的章节具体安排如下：

第一章，绪论。首先介绍了评分卡在金融领域发展的背景，分析评分系统与车联网的内在联系，其次对国内外评分卡和车联网领域的研究成果及现状进行分析，最后通过对国内外文献的学习，总结得出基于车联网数据构建驾驶安全评分模型的重要性，以此作为本篇文章研究工作的切入点。

第二章，预备知识。本章首先详细地介绍基于车辆网数据构建驾驶行为安全评分模型所需要的理论基础知识。以倒叙的方式，介绍为实现安全评分模型所需要的准备工作，而不断地引进数据准备和处理的理论方法。因此，本章选择详略得当、重点突出地对理论进行介绍。

第三章，数据预处理。在这一章节，首先介绍了数据的来源和基本情况，然后对数据进行清洗工作，进行描述性统计分析、相关性分析、驾驶风格聚类分析。特征工程主要包括对驾驶行为进行特征选择和 *woe* 分箱处理，完成构建评分模型所需数据的准备工作。

第四章，基于 Logistic 回归构建驾驶行为评分卡。本章的工作是建立在上一章的工作基础之上，构建不同标注下的评分卡模型。最后将评分结果进行对比分析，给出在实际操作中可以作为参考的评分结果。建议取判定有危险驾驶行为的阈值为两种模型阈值的均值。即当新数据的预测结果大于 0.525 时，预判为有危险驾驶行为。

第五章，基于 XGBoost 模型构建驾驶行为评分卡。考虑特征选择对驾驶安全评分卡模型建立的重要性，本章使用了 MIV 算法进行特征选择。于是本章先介绍 BP 神经网络

和基于 BP 神经网络的 MIV 算法，然后引进 XGBoost 模型，建立融合平均影响值特征选择的 XGBoost 驾驶行为安全评分模型。

第六章，总结。在这一章节回顾了全文的主要研究内容，同时进行合理的展望。

第二章　预备知识

本文通过对国内外文献的学习，研究分析评分卡模型在金融风险场景中的实际应用，发现在借贷风险控制模型中，主要分为三个阶段：分别为贷前、贷中、贷后三个阶段，其中贷前阶段是控制风险的关键。由于需求的驱动，相应的建模体系也就发展的越成熟，可应用的领域越广泛。因此，本文将贷前评分 A 卡简称评分卡 [14]，车联网数据的评分卡建模流程，如图 2.1。

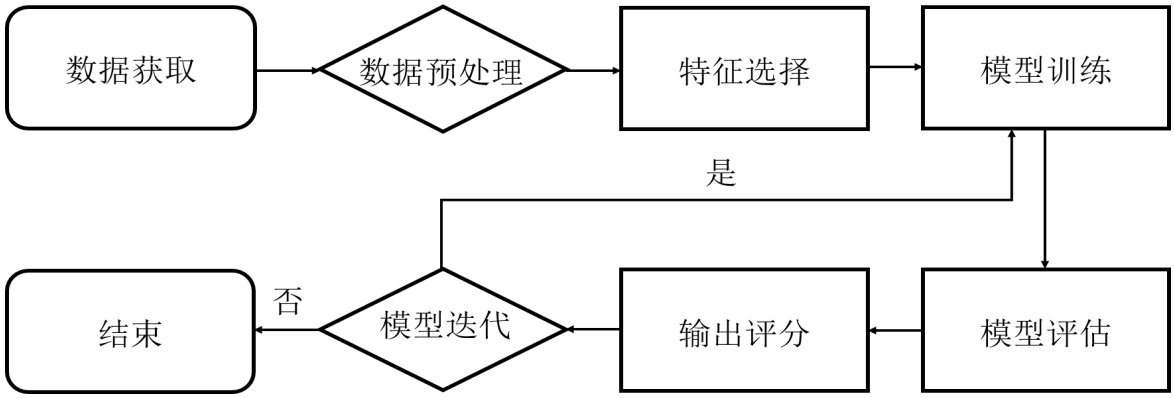

图 2.1 车联网数据的评分卡流程图

§2.1 评分卡理论

Logistic 回归 [35] 用于处理二分类数据，在评分卡模型中使用最为广泛。对驾驶行为数据进行监督学习，得到高显著性且符合实际意义的回归系数，预测驾驶行为的不安全概率，并将预测的概率转换为驾驶行为评分。与普通的分类模型不同的是，Logistic 回归通过增加 *sigmoid* 函数将预测值映射到 (0, 1) 区间。

$$g(z) = \frac{1}{1 + e^{-z}}.$$

其中 $z = \beta^T X$，X 为输入的 $n * m$ 特征矩阵，β 为 $n * 1$ 的系数矩阵，$g(z)$ 为 Logistic 回归输出概率值。

在车联网驾驶行为数据中，1 表示不安全驾驶行为，0 表示安全驾驶行为，模型预测安全的概率为：

$$P(Y = 0 \mid X;\beta) = g(z).$$

不安全的概率为：

$$P(Y = 1 \mid X;\beta) = 1 - g(z).$$

处理这类数据，通常根据实际生活的需要给 $g(z)$ 函数动态地设定阈值为 α，如果 $g(z) > \alpha$，则 $Y = 1$，否则，$g(z) \leq \alpha$，则 $Y = 0$。

通过 Logistic 回归预测驾驶行为的不安全概率，根据预测的概率，本文将安全驾驶和不安全驾驶行为的比值定义为 $Odds$ [13]：

$$Odds = \frac{p}{1-p}. \tag{2.1.1}$$

变形之后为：

$$p = \frac{Odds}{1 - Odds}.$$

定义评分卡与 $Odds$ 之间的映射关系，如式 (2.1.2)：

$$\begin{aligned} Score =& A - B\ log(Odds) \\ =& A - B\ log\frac{p}{1-p}. \end{aligned} \tag{2.1.2}$$

其中，A，B 为常数，令 $\theta_0 = \frac{p}{1-p}$ 定义：

$$\begin{aligned} P_0 &= A + B\ log(\theta_0) \\ P_0 + PDO &= A + B\ log(2\theta_0). \end{aligned} \tag{2.1.3}$$

将公式 (2.1.3) 代入 (2.1.2) 即可求解常数 A，B：

$$B = \frac{PDO}{log(2)}.$$

$$A = P_0 - B\,ln(\theta_0).$$

其中，P_0 是针对特定比率 $Odds$ 专门设定的分数，PDO 为比率 $Odds$ 翻倍时对应的分值。Logistic 回归输出预测概率 p 为式 (2.1.4) [13]：

$$p = \frac{1}{1 + e^{-\beta^T X}}. \tag{2.1.4}$$

其中，β 为 $n * 1$ 的系数矩阵。式 (2.1.4) 经过转换得到：

$$ln(\frac{p}{1-p}) = \beta^T X. \tag{2.1.5}$$

结合式 (2.1.1) 可得：

$$log(Odds) = log(\frac{p}{1-p}). \tag{2.1.6}$$

再结合式 (2.1.5)，(2.1.6) 有：

$$log(Odds) = \beta^T X = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n. \tag{2.1.7}$$

通过式 (2.1.2)，(2.1.7) 有：

$$\begin{aligned} Score &= A - B(\beta^T X) \\ &= A - B(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n) \\ &= (A - B\beta_0) - B\beta_1 x_1 - \cdots - B\beta_n x_n. \end{aligned} \tag{2.1.8}$$

其中，x_i 为 $n * 1$ 的列向量。

由于 Logistic 回归具有可解释强，适应范围广的特点。将特征进行分箱后计算 woe 值，然后将分箱后的特征转换成 woe 值，woe 的计算公式为：

$$woe_i = ln(\frac{bad_i}{bad_T} / \frac{good_i}{good_T}) = ln(\frac{bad_i}{bad_T}) - ln(\frac{good_i}{good_T}).$$

本文将车联网数据进行分箱处理，结合数据的特点进行灵活分箱，构建驾驶行为安全评分卡。应用式 (2.1.8) 给出特征分箱在驾驶行为评分中的具体转化形式：

$$\begin{aligned} Score =& A - B\left[\beta_0 + \beta_1(\delta_{11}\omega_{11} + \delta_{12}\omega_{12} + \cdots) + \beta_2(\delta_{21}\omega_{21} + \delta_{22}\omega_{22} + \cdots)\right. \\ &\left. + \beta_3(\delta_{31}\omega_{31} + \delta_{32}\omega_{32} + \cdots) + \cdots + \beta_n(\delta_{n1}\omega_{n1} + \delta_{n2}\omega_{n2} + \cdots) + \cdots\right]. \end{aligned}$$

其中，δ_{ij} 为示性函数，代表第 i 个特征的第 j 个分箱，ω_{ij} 代表第 i 个特征的第 j 个分箱的 *woe* 值。将 Logistic 回归模型转化评分的具体过程汇总为表 2.1。

表 2.1 评分卡

特征	箱数	分值	特征	箱数	分值	$\cdots$	特征	箱数	分值
基准点	-	$A - B\beta_0$	基准点	-	$A - B\beta_0$	$\cdots$	基准点	-	$A - B\beta_0$
	1	$-B\beta_1\omega_{11}$		1	$-B\beta_2\omega_{21}$			1	$-B\beta_n\omega_{n1}$
	2	$-B\beta_1\omega_{12}$		2	$-B\beta_2\omega_{22}$			2	$-B\beta_n\omega_{n2}$
x_1	$\cdots$	$\cdots$	x_2	$\cdots$	$\cdots$	$\cdots$	x_n	$\cdots$	$\cdots$
	$\cdots$	$\cdots$		$\cdots$	$\cdots$	$\cdots$		$\cdots$	$\cdots$
	k_1	$-B\beta_1\omega_{1k_1}$		k_2	$-B\beta_2\omega_{2k_2}$			k_n	$-B\beta_n\omega_{nk_n}$

针对如何判断模型在训练集和测试集上表现的好坏程度，下面则给出 AUC (Area Under Curve)和 KS (Kolmogorov-smirnov) 两个指标的定义。

首先，给出混淆矩阵 [35]，有：

表 2.2 混淆矩阵

	是(危险)	否(安全)
是(危险)	TP(True Positive)	FN(False Negative)
否(安全)	FN(False Negative)	TN(True Negative)

然后，根据表 2.2 计算得到：

$$Precision = \frac{TP}{TP + FP}.$$

$$Recall(TPR) = \frac{TP}{TP + FN}.$$

$$FPR = 1 - \frac{FP}{TN + FP}.$$

其中，FPR 为假正率 (False Positive Rate)，TPR 为真正率为 (True Positive Rate)，以 TPR 为纵轴，FPR 横轴绘制的曲线即 ROC 曲线。曲线下方的面积大小代表 AUC 值，AUC 的值越大，说明模型整体训练的效果越好。定义 TPR 和 FPR 差的最大值为 KS 的值，即 $max(TPR - FPR)$。以 TPR，FPR 为两条纵轴，0 到 1 之间的概率值为横轴绘制曲线，即得到 KS 曲线，KS 值的计算公式为：

$$\text{KS} = max(TPR - FPR), \quad KS \in (0,\ 1).$$

§2.2 Kmeans 算法

1967 年 MacQueen [35] 提出的 Kmeans 算法，因为其简单性和优秀的算法性能，目前仍然是使用最多的聚类算法。作为一种无监督的机器学习方法，Kmeans 使用无标签的数据集进行迭代训练，划分为 K 个簇，其主要步骤分为聚类中心更新和样本的簇分配步骤。

Kmeans 算法分为以下几步：

1) 给定聚类簇数 k，随机初始化 k 个聚类中心：

$$\mu^0 = \{\mu_1^0,\ \mu_2^0,\ \cdots, \mu_k^0\}.$$

2) 设最大迭代次数为 T，样本数为 N，在第 $t = 0, 1, 2, ..., N$ 次迭代时，分别计算数

据集中每个样本点 x_i 与每个聚类中心 μ_k^{t-1} 的距离。将每个样本点分配给与其距离最小的簇。

3) 遍历每个簇 C_j，$1 \leq j \leq K$，统计簇 C_j 中的数据信息，设置新的聚类中心 μ_j^t 为簇内数据的均值：

$$\mu_j^t = \frac{1}{|x_i \in C_j|} \sum_{x \in C_j} x_i, 1 \leq j \leq K.$$

于是得到新的聚类中心：

$$\mu^t = \{\mu_1^t, \mu_2^t, \cdots, \mu_K^t\}.$$

4) 重复 2) - 3)，当聚类中心 μ^t 不再更新，或者达到了最大的迭代次数 T，得到最终划分为 k 个簇的数据集。

上述算法性能的关键在于第 2) 步中距离度量的选择，设两个数据点 x_i 和 x_j 表示为：$x_i = (x_{i1}, x_{i2}, ..., x_{im})$，$x_j = (x_{j1}, x_{j2}, ..., x_{jm})$，$m$ 为特征维度。度量方式通常有以下几种：

1. 欧式距离

欧式距离 [36] (Euclidean Distance) 表示在 m 维空间中，两点之间的绝对距离，表示为：

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}.$$

2. 曼哈顿距离

曼哈顿距离 (Manhattan Distance) [37] 也被称为出租车距离。

$$d_{ij} = |\sum_{k=1}^{m} (x_{ik} - x_{jk})|.$$

3. 马氏距离

马氏距离 [38] (Mahalanobis Distance) 用来表示数据的协方差距离，能够计算两个数

据相似度的方法：

$$d_{ij} = \sqrt{(x_i - x_j)S^{-1}((x_i - x_j)}.$$

其中，S 为数据的协方差矩阵。

§2.3 随机森林算法

1996 年 Breiman [39] 提出了 Bagging 算法，Random Forest (RF) [40] 是一种继承了 Bagging 算法原理的思想，由 m_{tree} 个决策树组成的学习方法，其中每棵树都是通过自举构建的最优分类器。该方法具有处理数百个输入变量的优点，不用进行数据降维预处理步骤，并且不太容易过拟合 [41]。此外，该方法还可以对分类的特征重要性进行排序。在随机森林分类器中设置了两个关键参数：决策树的数量 n_{tree} 和每个节点的特征数量 m_{try}。

随机森林算法分为以下几步：

1) 基于 boostrap 抽样方法有放回从原数据中抽取 n 个样本集，来构建 n_{tree} 棵决策树，由 bootstrap 随机抽样的样本，每次抽取都有未被抽中的样本，这些未被抽中的样本集组成 n 个袋外数据集。

2) 假设有 k 个特征，随机森林随机从这 k 个特征中抽取 m_{try} 个特征 ($m_{try} \leq$ k)，然后通过这 m_{try} 个特征的信息量确定决策树节点的最优切分。其中每颗决策树不做任何裁剪。

3) 多颗决策树共同组成了随机森林，当有新的数据到来时，使用随机森林的每颗决策树进行分类，得到分类票数最多的标签为分类结果。

由于第 1) 步采用的是 boostrap 进行随机采样构建每颗决策树，这种方法决定了每颗决策树的随机性，提升了分类器的多样性，这样就不会因为过多相似的分类器使得分类的错误率提升。随机森林根据这些分类器进行预测，具有更好的泛化性。

§2.4 Boruta 算法

Boruta 包 [42] 使用一个随机森林包装器，广泛地用于特征选择。其通过原始特征随机打乱的值 (被称为 shadow 特征) 和原始特征进行混合来计算特征的重要性。Boruta 算法是围绕随机森林分类算法构建的包装器。随机森林基于 Z-score 给出重要度得分，但是单独的 Z-score 不能准确地得到特征的相关性，因此需要一些其他标准来区分因变量中重要和不重要特征。而 Boruta 算法 [43] 可以得到数据集中存在的所有重要且与输出变量相关的特征。Boruta 算法分为以下几步：

1) 创建 shadow 特征，这些特征由原始特征进行随机打乱得到，将这些特征添加到数据集中。

2) shadow 特征和原始特征进行混合，共同训练模型。然后计算特征重要性度量 (通常使用平均降低精度)，得到每个特征的重要性。值越大，其重要性程度越高。

3) 评估 Z-socre。在每一次迭代中，Z-score 会检查真实的特征是否比最好的 shadow 特征更重要。设最好的 shadow 特征的 Z-score 为 S_{max}。如果真实特征的 Z-score 显著高于 S_{max}，标记其是重要的特征。对不确定重要性的特征执行双侧检验。如果 Z-score 显著低于 S_{max}，标记其为不重要，删除这个特征，防止降低模型的性能。

4) 重复2) - 3)，直到所有的特征都已被标记，或者达到了最大的迭代次数。

第三章　驾驶行为数据预处理

§3.1　数据介绍与处理

§3.1.1　数据介绍

JSON (JavaScript Object Notation) 是一种易于机器存储、编写、解析、生成的新型数据交换格式。本章所使用的数据主要来自某市的车联网数据，数量为 60 辆，时间为 2021 年 3 月，记录方式为按天拆分的 JSON 格式数据。每辆车均收集到大量的数据信息，以键值对的形式存储。数据经过合并整理，共计 26 个驾驶特征，250 多万条驾驶数据。结合联网存在信号线路问题，发现有部分数据和特征存在缺失、无效采集、部分异常等现象。下面将结合数据特点和生活实际进行分析处理，先给数据查阅简表 3.1。

表 3.1 特征简表

gps_lat	纬度	elec_park_status	电子驻车状态
gps_lon	经度	first_safebelt_staus	驾驶员安全带状态
last_oil	剩余油量值	mainten_mileage	剩余保养里程
total_mileage	总里程	gear_info	档位信息
batt_power	蓄电池电量	brake_pedal_switch	制动踏板开关
car_speed	车速	gps_direction	GPS方向
engine_status	发动机状态	gps_time_stamp	GPS 发生时间
engine_speed	发动机转速	env_temperature	环境温度
remain_mileage	剩余续航里程	incar_temperature	发动机水温
soc_warn	动力电池 SOC 跳变报警	engine_water_temperature	发动机水温
right_light_status	右转向灯开关状态	gear_oil_temperature	变数器油温
safeball_status	安全气囊碰撞状态	safeball_signal	安全气囊碰撞信号
left_light_status	左转向灯开关状态	second_safebelt_staus	副驾安全带状态

§3.1.2 描述性统计分析

面对大量需要清洗的数据，描述性统计分析能够快速掌握数据的整体情况，检测出可能的异常值。通过柱状图、箱线图、密度曲线等可视化手段给出数据频率和占比、离散程度、缺失程度、大致分布等重要指标。对数据进行简明扼要的分析与处理，例如，缺失程度高于 50% 的特征将会被剔除；缺失程度中等的特征一般会采用均值和众数进行数据回填；轻度缺失的特征会删除缺失部分的数据保留特征。数据处理后，得到缺失程度超过 50% 的特征有 soc_warn、safeball_status、safeball_signal 这 3 个特征。由于车联网信号问题，导致数据上传错误共有 4 个特征。存在离群值共计 7 个特征。数据特征表现良好的特征这里不做赘述，上述各具体特征详细处理方式，由表 3.2 给出。

表 3.2 特征预处理

特征状态	特征	百分比	处理方式
缺失值	soc_warn	100%	1
	safeball_status	100%	1
	safeball_signal	100%	1
存在异常值	incar_temperature	100%	1
	env_temperature	> 60%	1
	charg_conn_status	100%	1
	batt_power_unit	100%	1
存在离群值	gps_lat	< 1%	2
	gps_lat	< 1%	2
	last_oil	< 1%	2
	total_mileage	< 1%	2
	car_speed	< 1%	0
	engine_speed	< 1%	0
	engine_water_temperature	< 1%	0
	gear_oil_temperature	< 1%	0

其中，0 代表删除数据；1 代表删除特征；2 代表不做处理。

如表 3.2 所示，本文给出数据异常和存在离群值两种情况，三种处理方式，分别举一个特征为例加以说明。

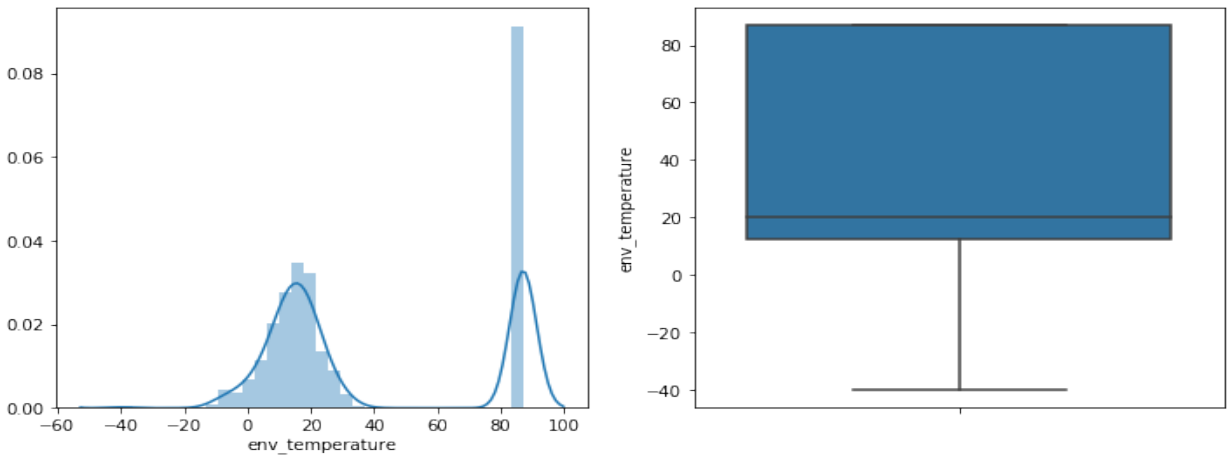

图 3.1 环境温度分布图

如图 3.1，得到环境温度 (env_temperature) 的最大值为 87 ℃，最小值为 - 40 ℃，均值高于 39℃ 且超过半数的数据均高于均值。由常识可知，这样的环境温度是不可能存在的。本文获取数据的时间是 2021 年 3 月份，某市 3 月份的天气还比较寒冷，平均气温也仅能达到 10℃ 左右，而箱线图也显示绝大多数温度值均在 20℃ 以上。另外，所有驾驶数据均来自同一个月份，驾驶员所处的环境温度因素基本是保持一致的。因此，像这样与常识相悖且意义不大的特征本文选择将该特征剔除。

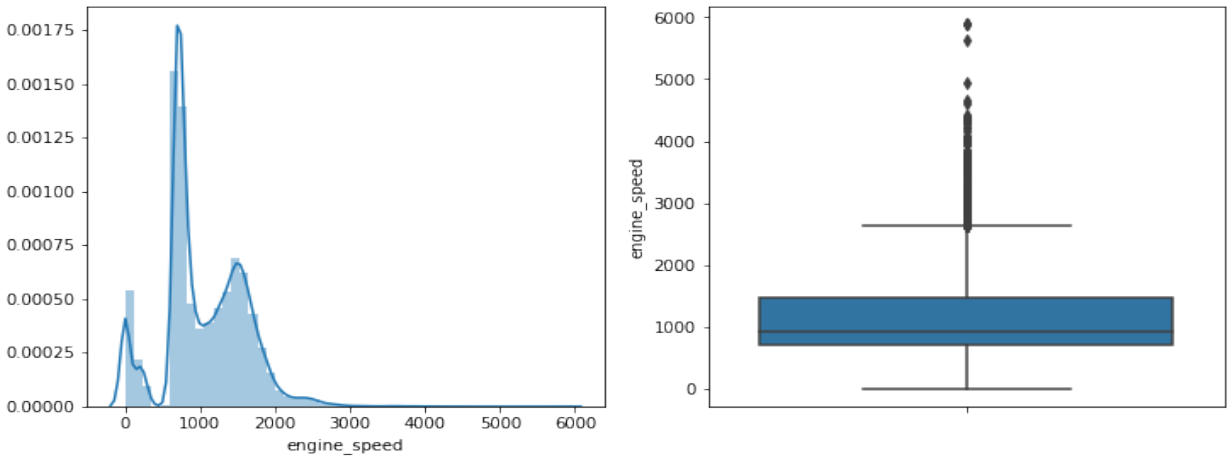

图 3.2 引擎速度分布图

对于引擎速度 (engine_speed) 这个特征，对于有过驾驶经验的司机来说，汽车转速在 1000 - 3500 r/min 属于是正常值范围。一般在行驶过程中会将转速保持在 2200 - 3500 r/min 左右。而收集的数据最小值为 0 r/min，最大值为 5882 r/min，均值为 1051.36 r/min，与转速标准对比，驾驶行为数据看似是不合理的。正因为收集到的数据有一部分不在这个标准里，才说明这个驾驶行为里面存在安全隐患，是由于驾驶员操作不当造成的，十分具有研究价值。因此，图 3.2 中箱线图所显示的离群值，本文没有足够的理由说明它存在异常，故不做处理，选择保留。

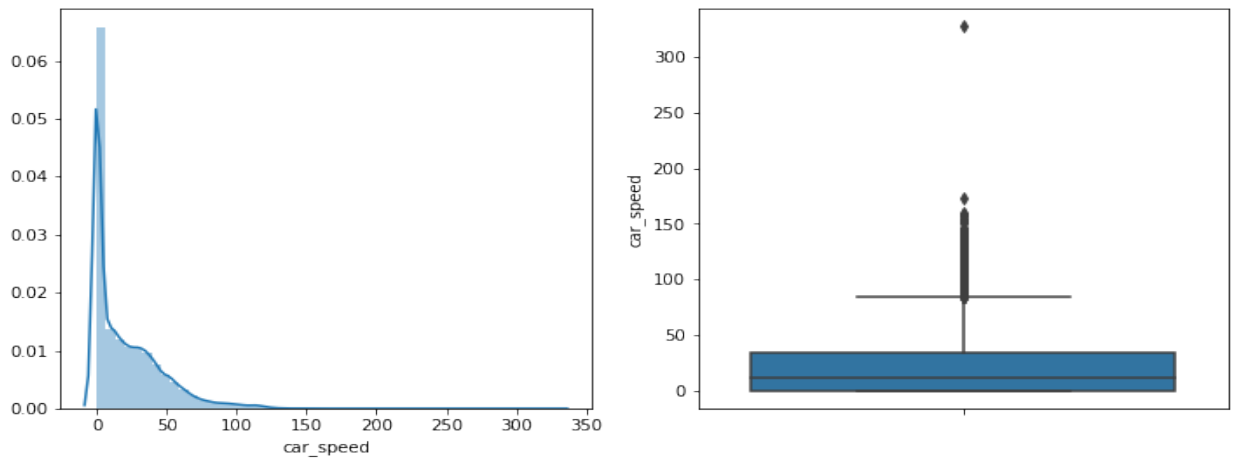

图 3.3 车速分布图

对于车速 (car_speed) 这个特征，在图 3.3 中，离群值显示速度的值为 327 km/h，这是不可能的。正常行驶的车数超过 120 km/h 已经非常危险，而大部分数据由箱线图显示均在一个较为合理的范围。因此，剔除掉这类离群异常的数据点。

以上通过描述性统计分析，剔除了缺失程度达到 100% 的 3 个特征，由于车联网信号问题，剔除出现数据异常共 4 个特征，进行异常值处理共计 7 个特征。

§3.2 相关性分析

Pearson 提出相关系数，之后 Spearman 又提出等级相关系数。前者是检验变量之间的相似程度，后者是检验变量之间是否存在多重共线性。经过上一小节的描述性统计分析之后，保留了 19 个特征。将其相关性检验结果进行分析得到，特征之间存在一定的相关关

系。其中右转向灯状态 (right_light_status) 和左转向灯状态 (left_light_status) 高度相关，车速 (car_speed) 和引擎速度 (engine_speed)，引擎水温 (engine_water_temperature) 和齿轮油温 (gear_oil_temperature) 显著相关。本文怀疑是因为特征的语义比较相近而导致相关性系数偏高，但是驾驶员安全带状态 (first_safebelt_staus) 和副驾驶安全带状态 (second_safebelt_staus) 却没有因为功能和语义相似而存在较高的相关系数。这说明用相关系数评价特征之间的相关性是具有一定的可信度，能够区分特征间的不同。

表 3.3 相关性判定准则

相关性	微弱相关	低度相关	显著相关	高度相关
相关系数取值范围	$0.0 < \|\gamma\| < 0.3$	$0.3 \leq \|\gamma\| < 0.5$	$0.5 \leq \|\gamma\| < 0.8$	$0.8 \leq \|\gamma\| < 1.0$

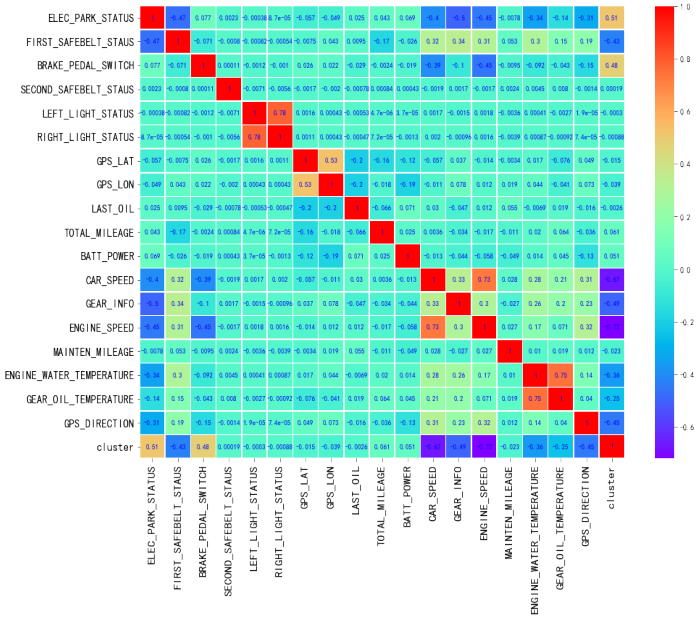

图 3.4 相关性检验结果

对于相关系数超过 0.8 的驾驶行为，会选择删除处理，但是经过 Pearson 相关系数检验得知[44]，不同驾驶行为特征之间的相关系数均不大于 0.8，可以初步判断驾驶行为特征之间不存在多重共线性。一般情况下，对于弱和强的相关规定，如表 3.3 所示。其中，具有较高相关系数的驾驶行为本文要进行多重共线性检验，因为相关系数高，并不意味着就具有多重共线性。

由图 3.4 可知驾驶行为特征间的具体相关关系。经过 Spearman 秩和检验 [45]，得到的结论是右转向灯开关状态 (right_light_status) 和左转向灯开关状态 (left _light_status) 以及车速 (car_speed) 和引擎速度 (engine_speed)，这两对特征之间存在共线性。而引擎水温 (engine_water_temperature)和齿轮油温 (gear_oil_temperature) 之间不存在共线性。事实上，结合所掌握的知识经验可以知道，车速和引擎速度之间是后者影响着前者，前者制约后者。如果已知轮胎直径，主减速比，对应挡位传动比这三个参数，可以利用下面的公式计算出轮胎的直径，以及车速与引擎速度之间的线性关系。

$$D = 2W \cdot AR + 25.4\, d. \tag{3.2.1}$$

其中 D 为轮胎直径, W 为断面宽度，AR 为扁平比，d 为轮钢直径。

计算出轮胎直径后，再将发动机转速或者车速值带入式 (3.2.1) 有：

$$V = \frac{60\,\pi \cdot D \cdot V_{engine}}{1000\, k_1 \cdot k_2}.$$

其中，V 为车速，V_{engine} 为发动机转速，k_1 为主减速比，k_2 为对应挡位传动比。

计算出发动机转速或者车速的值，通过这个结果和现有数据进行对比。可以知道，造成车速和引擎速度这两个驾驶行为特征没有高度共线性的原因，是因为驾驶员在驾驶过程中，所处驾车环境是复杂多变的。因此真实值和理论值存在偏差是合理的，但二者存在共线性。

§3.3 驾驶风格的标注与分析

§3.3.1 驾驶风格聚类分析

由于处理过的数据仍处于无标注状态，可以使用机器学习方法中的 Kmeans 聚类方法对该数据进行处理。在聚类之前先对各驾驶行为特征做标准化变换，根据数据的特征将拥有相似驾驶风格的驾驶行为聚为一类。从 sklearn.cluster 导入 Kmeans 包，设置关键参数进行调参。首先设置迭代的最大次数为 300，防止类中心选取不合适，聚类算法进入死循环。设置算法的结束条件为平方误差和均方误差都满足给定误差。最后输出类中心，实际的迭代次数，重要特征得分，每一类的样本数。

聚类的目的就是把驾驶行为特征按照规则划分为几类，而这个类的个数是根据数据的特征和距离类中心的平均距离来确定的。第一类有 123,0510 条驾数据，第二类有 68,1577 条数据。根据聚类中心的特点，构建驾驶安全评分模型。对类中心进行方差齐次检验，如果验证得到组间差异性显著，同时组内几乎无差异性，就说明分组效果良好，组内驾驶员驾驶风格有一定的共性。如果满足这个假设就进一步说明 Kmeans 聚类分组是可靠的。

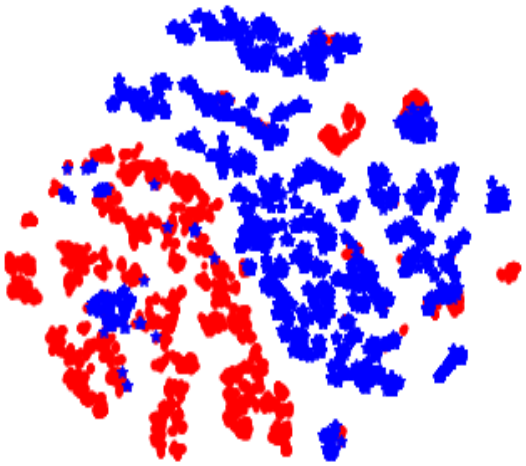

图 3.5 聚类结果降维展示图

为了使聚类结果可视化，本文选择从 sklearn.manifold 导入 t-SNE 包 (t-distributed Stochastic Neighbor Embedding)，通过该方法将高维空间的数据点映射到低维空间中，从而达到聚类降维可视化的目的。根据 Kmeans 的聚类结果，得到数据的二维嵌入，如图 3.5 所示，为降维后数据的可视化结果。根据红色和蓝色数据点的分布情况，可以直观地得到驾驶行为被分成两类且区分度显著。

分析类中心的特点可知，第一类驾驶员在车速 (car_speed)、档位信息 (gear_info)、GPS 方向 (gps_direction)、驾驶员安全带状态 (first_safebelt_staus) 等特征上得分较高，在行驶过程中更在意对速度的掌控和对方向盘的把控，更多的注意力集中在驾驶操作行为上，说明在行驶过程中驾驶员为速度耗油操作型。第二类驾驶员在电子制动开关 (elec_park_status)、制动踏板开关 (brake_pedal_switch)、蓄电池电量 (batt_power)、副驾驶安全带状态 (second_safebelt_staus) 等特征得分较高。主要表现在能够很好地关注仪表盘上各项指标并兼顾副驾驶的安全状态，没有把更多的注意放在车速和操作上，很好的关注了汽车本身

的性能指标，说明在驾驶过程中驾驶员为稳定熟练型。

对于第一类驾驶员给出的建议是：要注意行车速度的控制，关注仪表盘，例如发动机水温，变速器油温，剩余油量值等各项指标，避免由于车速变化较快使发动机水温，变速器油温升高，使发动机不能得到很好的冷却，造成不可逆的损害。第二类驾驶员给出的建议是：相对于第一类驾驶员，具有较好的驾驶习惯，能够兼顾副驾驶安全带的状态，关注各项仪表盘指标的同时也要避免急刹车，急转弯等危险行为的发生。

§3.3.2 驾驶风格方差齐性分析

本文在这一小节之前，已经完成了聚类分析，通过调参并结合算法的收敛性，将驾驶员行为记录划分为两类。在这一小节对驾驶员驾驶行为记录分类的合理性进行检验，利用的方法为基于正态样本的方差齐性检验。

首先，从 scipy 中导入 stats 包，将处理过的驾驶行为聚类中心进行正态性检验，得到的结果如表 3.4 所示。在显著性 0.05 的水平下，P 值均大于 0.05，否定了两种驾驶员行为记录分类方差不一致的情况，说明两个类别来自同一个正态总体，即进一步说明这两个类别的驾驶行为记录均服从正态分布。然后，继续从 scipy.stats 中导入基于正态样本方差齐性检验的 levene 方法，得到的结果如表 3.4 最后一列所示，检验结果均大于 0.05，说明划分的驾驶行为记录这两个类别方差波动具有一致性，即来自同一总体的两个样本方差具有非齐性。这样两个类别的样本才具有可比性。通过比较这两个类别的均值，发现它们不同如表 3.4 第二列所示，这使得本文有理由认为驾驶行为记录划分为两类是合理的，而且从检验的结果可以看出，两类样本涵盖的驾驶信息重合性很低。同时检验结果和聚类结果降维展示图 3.5 所示保持一致。

表 3.4 驾驶风格方差齐性检验

组别	均值	方差	统计量	P 值	齐性检验结果
第一组	0.09607	0.24990	0.21628	0.32210	3.03426
第二组	-0.17906	0.46589	0.21630	0.32211	0.09057

为了进一步说明上述分类的合理性，将驾驶记录为分类为三类及以上，并进行相应的检验，未取得合理的检验结果。由于数据涵盖驾驶行为特征数量有限，如果强行划分

为两类以上，就会出现类别间信息量交叉，不具有类别显著的独有特性。因此，本文选择将驾驶行为记录划分为两类。

§3.4 特征工程

§3.4.1 驾驶行为特征选择

经过前面几节对数据的分析与处理，最后保留了 19 个特征。下面为了进一步地提高模型的性能，主要通过随机森林、向前逐步回归和 Boruta 算法进行特征选择。

表 3.5 特征选择结果

RF	Forward	Boruta
car_speed	gear_info	gps_lat
gear_info	brake_pedal_switch	gps_lon
engine_water_temperature	gps_direction	last_oil
brake_pedal_switch	car_speed	total_mileage
first_safebelt_staus	engine_water_temperature	batt_power
gear_oil_temperature	first_safebelt_staus	car_speed
elec_park_status	gear_oil_temperature	gear_info
gps_lat	elec_park_status	elec_park_status
batt_power	total_mileage	first_safebelt_staus
total_mileage	gps_lat	mainten_mileage
gps_lon	gps_lon	brake_pedal_switch
mainten_mileage	last_oil	engine_water_temperature
last_oil	mainten_mileage	gear_oil_temperature
	batt_power	

三种特征选择方法得到的结果如下所述：

随机森林在特征选择的过程中，通过计算驾驶行为的贡献程度，剔除掉贡献值为 0 的特征，在表 3.5 中按贡献程度从大到小的顺序给出。

向前逐步回归在进行特征选择的时候，通过不断剔除回归系数不显著的驾驶行为，并不断地引入新的驾驶行为，直到所有的系数都通过显著性检验。另外本文还对驾驶行为增加了确保 *Aic* 最小的限制，在表 3.5 中，按照 *Aic* 的值从小到大的顺序给出。

在 Boruta 方法中，驾驶行为的排序越小贡献越大。于是本文选择排序为 1 的驾驶行为，并在表 3.5 中给出。

最终三种方法筛选出的驾驶行为个数为 13，14，13。特别的是，Boruta 算法和随机森林特征选择的结果相同，而向逐步回归的结果，比这两种方法得出的结果增加了一个 GPS 方向的特征。最后本文选取三种方法特征选择结果的并集，将此结果作为构建评分模型分箱的特征，将特征选择的结果汇总为表 3.5。

§3.4.2 驾驶行为分箱

根据 §3.4.1 小节的特征选择结果，得知特征选择的结果有 14 个驾驶行为被保留下来。基于 Logistic 回归构建驾驶行为安全评分模型之前，要对数据进行 *woe* 分箱处理。于是需要对入模的驾驶行为特征增添了两个限制条件，即 IV 值要大于 0.03 和 *woe* 分箱中坏样本个数要保持单调。表 3.6 给出了 IV 值的选取规则。

表 3.6 IV 值的贡献能力

IV	贡献能力
$[0.0, 0.02)$	无贡献能力
$[0.02, 0.1)$	弱贡献能力
$[0.02, 0.03)$	微弱贡献能力
$[0.03, 0.1)$	较弱贡献能力
$[0.1, 0.3)$	中等贡献能力
$[0.1, inf)$	强贡献能力

由表 3.6 可知 IV 值的大小反应的是对预测的贡献能力。当根据 IV 值选取入模特征时，要同时结合 *woe* 分箱的结果，灵活地进行选择。即根据驾驶特征在每一个箱的具体表现，适当地降低或者提高被选入模型的驾驶行为特征的 IV 值。本文考虑到 *woe* 分箱要

满足坏样本数量保持单调的要求，将入模的驾驶特征贡献能力的阈值提高到 0.1 的水平，相当于只有贡献能力达到 0.1 才有机会被选入驾驶安全评分模型中进行训练，其余贡献能力未达到 0.1 的驾驶行为将被逐一剔除。表 3.7 和表 3.8 给出两例分组后 *woe* 值没有保持单调的驾驶特征作为展示，并分析被剔除的原因。

表 3.7 GPS 经度分箱结果

特征	分箱	箱内样本数	安全样本数	不安全样本占比	woe 值	分箱的 IV 值
gps_lon	(-inf,103.0)	70224	32324	0.53970	0.359940	0.012843
gps_lon	[103.0,112.0)	163510	59475	0.636261	0.040090	0.000356
gps_lon	[112.0,115.0)	111325	35446	0.681599	0.242048	0.008580
gps_lon	[115.0,116.0)	140598	63667	0.547170	-0.329839	0.021546
gps_lon	[116.0,inf)	248345	82920	0.666110	0.171560	0.009723

表 3.8 GPS 纬度分箱结果

特征	分箱	箱内样本数	安全样本数	不安全样本占比	woe 值	分箱的 IV 值
gps_lat	(-inf,24.0)	75629	23716	0.686417	0.264338	0.006927
gps_lat	[24.0,27.0)	116638	50236	0.569300	-0.240086	0.009401
gps_lat	[27.0,31.0)	129421	46222	0.642855	0.068698	0.000825
gps_lat	[31.0,33.0)	108644	47539	0.562433	-0.268038	0.010941
gps_lat	[33.0,37.0)	98110	34442	0.648945	0.095324	0.001199
gps_lat	[37.0,39.0)	47003	20763	0.558262	-0.284969	0.005358
gps_lat	[39.0,40.0)	64805	19982	0.691660	0.288808	0.007057
gps_lat	[40.0,inf)	93752	30932	0.670066	0.189401	0.004461

对于多次分箱后，GPS 经度和 GPS 纬度这两个特征的 *woe* 值仍没有处于单调的状态，这样的特征代入评分模型中，会导致模型的可解释变差，精确度不高。因此，这类特